The Future of Survey Research: Challenges and Opportunities

A Report to the National Science Foundation
Based on Two Conferences Held on
October 3-4 and November 8-9, 2012

Report Presented By

The National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences Subcommittee on Advancing SBE Survey Research:

Jon A. Krosnick – *Stanford University (chair)*
Stanley Presser – *University of Maryland*
Kaye Husbands Fealing – *University of Minnesota*
Steven Ruggles – *University of Minnesota*

Project Coordinator:

David L. Vannette – *Stanford University*

When the National Science Foundation Advisory Committee for the Social, Behavioral and Economic Sciences formed the subcommittee on Advancing SBE Survey Research, Janet Harkness was a member. We dedicate this report to her memory, to recognize her contribution to this subcommittee and the survey research community and the work of the National Science Foundation.

May, 2015

Table of Contents

Overview

For more than thirty years, the National Science Foundation (NSF) has supported data for research on a wide variety of topics by making awards to three major long-term survey efforts, the American National Elections Studies (ANES), the Panel Study of Income Dynamics (PSID), and the General Social Survey (GSS). In February 2012, the Advisory Committee for the Social, Behavioral, and Economic Sciences (SBE) was asked to provide advice about future investments in these surveys and others. The Advisory Committee then charged a subcommittee to provide that advice. The Subcommittee on Advancing SBE Survey Research is comprised of Jon Krosnick (Stanford University, chair), Janet Harkness (University of Nebraska, deceased), Kaye Husbands-Fealing (University of Minnesota), Stanley Presser (University of Maryland), and Steven Ruggles (University of Minnesota).

The Subcommittee submits this report to the Assistant Director of the SBE Directorate, with the purpose of providing advice related to how the Foundation can best use its resources to support research through survey data collection. Specifically, the report addresses the following questions, as requested:

1. What are the challenges facing survey-based data collection today (e.g., falling participation rates, rising costs, or coverage of frames)?

2. What innovations in survey methodology have taken place or are on the horizon?

3. How should SBE think about survey data in the context of the explosion of new digital sources of data? Does the Subcommittee see opportunities for blending data or mixed source methods that integrate existing administrative, commercial, or social media data with existing surveys to answer social science questions?

4. Given current challenges faced by survey research as well as the potential opportunities presented by new approaches to survey research, what types of questions will we be able to address with surveys in the future?

5. What is an overarching strategy for data collection that the Directorate supports (including, but not limited, to the three existing surveys), which could be used to guide planning for NSF-supported data in the future? This might include shared infrastructure across surveys, but should not be limited to that approach.

The report addresses the first four questions – which are about the current and future status of survey research in general (as opposed to uniquely about NSF funded surveys) – by drawing on the results of presentations we commissioned from leading experts at two conferences held at NSF in the fall of 2012. An extensive summary of the conferences is included as an appendix to this report. The fifth item in our charge – the strategic vision for SBE regarding the Big Three NSF surveys – was developed by the Subcommittee based on the proceedings of the conference and our own deliberations.

The two conferences we convened in October 2012 brought together leading scholarly experts on topics that fit into four broad areas. First, discussion on challenges faced in conventional survey research covered a broad landscape, including key topics such as: probability versus non-probability sampling methods; multi-mode survey techniques; optimizing response rates and how nonresponse affects survey accuracy; use of incentives in survey collection; survey design, visual displays and cognitive evaluation of survey instruments; proxy reporting; interviewing techniques and challenges; confidentiality, respondent attrition and data attrition; and computation of survey weights.

The second category of exploration focused on opportunities to expand data collection, including: paradata; the use of leave-behind measurement supplements and biomarkers; and specialized tools for measuring past events. Third, several methods of linking survey data with external sources were discussed, specifically: improving government, academic and industry data-sharing opportunities; linking survey data to official government records or with the Catalist Commercial Database; linking knowledge networks web panel data with external data; and the use of election administration data with other datasets. Lastly, discussion turned to improving research transparency and data dissemination, with a focus on: data curation; evaluating the usability of survey project websites; and the broader topic of the credibility of survey-based social science. Throughout the proceedings, conference participants explored steps that can be taken to enhance the value of survey methodology to a wide range of users, in academia, government, and the private sector.

Collectively, the conferences yielded several useful outcomes, including: (1) insights about how surveys should be done today to maximize data quality (thereby specifying how major infrastructure surveys should be designed and carried out), (2) important challenges facing the methodology, (3) best practices in data dissemination and data collection procedure documentation, (4) approaches that would be most desirable for large-scale infrastructure surveys to implement, and (5) research questions that merit future investigation.

Our report is organized as follows. Section 1 summarizes best practices for survey research recommended by the conference participants that should be useful to NSF programs as a touchstone in evaluating research proposals involving the collection of survey data. Section 2 offers a set of recommendations in response to the charge's final (fifth) question about an overarching strategy for NSF's major infrastructure projects that involve the collection of survey data. The Appendix is a rich compilation of detailed summaries of presentations at the two conferences sponsored for this study. (For transcripts of each presentation and the visual displays employed during the conferences see https://iriss.stanford.edu/content/future-survey-research-nsf.)

Section 1 – Best Practices for Survey Research

In this section, we present an extensive series of recommendations for best practices based on the conference proceedings.

Probability vs. Non-probability Sampling

1. To make an inference from a sample to a population requires assumptions. These assumptions are well understood only for samples drawn with known or calculable probabilities of selection. Thus, only such samples are widely accepted (Baker et al., 2013).

Response Rates

1. Researchers should carefully document nonresponse using the American Association for Public Opinion Research (AAPOR) categories and make efforts to understand the correlates of nonresponse so that users of the data are alerted to potential nonresponse bias (Groves and Couper, 2012).
2. Nonresponse bias is rarely notably related to nonresponse rate, so reducing nonresponse bias should usually be a more important goal than maximizing response rates (Groves and Peytcheva, 2008).
3. Refusal conversion is critical for increasing response rates and also may affect reductions in nonresponse bias (Groves and Heeringa, 2006).
4. Effective training of interviewers and other survey staff has been demonstrated to increase response rates (Conrad et al., 2012; Durrant et al., 2010).
5. Careful design of survey materials influences response rates. These materials include invitations, self-administered questionnaires, and other items that respondents see (Vicente and Reis, 2010).
6. Incentives, particularly pre-paid monetary incentives, can notably increase response rates (Singer et al., 1999).
7. Multiple contact attempts may increase response rates substantially (Keeter et al., 2000).
8. Advance letters notifying household members of their invitation to participate may increase response rates (Link and Mokdad, 2005).
9. To best understand and minimize nonresponse bias, response rates should be modeled as a stochastic process based on the association between response propensity and the characteristic being estimated (Brick, 2013).

Data Collection Modes (e.g. telephone, web, face-to-face, etc.)

1. Response rates and data quality are often highest with face-to-face surveys and lower in other modes, such as telephone interviewing, paper questionnaires, and Internet data collection (Hox and de Leeuw, 1994).
2. To minimize non-coverage in telephone surveys, both cellphones and landlines should be called (Brick et al., 2007).
3. Collecting data via multiple modes is an effective way to reduce costs and increase response rates (de Leeuw, 2005).

4. To maximize comparability between modes in a mixed-mode design, a unimode design for the questionnaire can be implemented, avoiding design features that are not replicable in every mode used (Dillman and Christian, 2005).
5. To minimize total error in mixed-mode surveys, use a responsive design framework – taking advantage of every measurement error reduction affordance available in each mode, even if not replicable between modes (Groves and Heeringa, 2006).

Survey Incentives – Paying Respondents to Participate

1. To increase response rates, use pre-paid incentives (Church, 1993).
2. In interviewer-mediated surveys, promised incentives typically have no effect on response rates (Singer, 2002).
3. Additional incentives paid (differentially) to recruit people who initially refused to become respondents may increase response rates among these reluctant people and reduce nonresponse bias (Singer, 2002; Singer et al., 1999).

Using Responses from Proxies (e.g., Household Members Other than the Respondent)

1. Proxy reports can be used to reduce costs in face-to-face interviewing (Boehm, 1989).
2. Proxy reports can best be used to measure observable information about the target respondent, rather than unobservable information such as attitudes (Cobb and Krosnick, 2009).

Improving Question Design to Maximize Reliability and Validity

1. To reduce satisficing behavior, minimize task difficulty and maximize respondent motivation (Krosnick, 1999).
2. To reduce respondent frustration, questionnaire design should follow conversational norms to the greatest extent possible, e.g., via adherence to the Gricean Maxims (Grice, 1975).
3. Optimal approaches to questionnaire design include:
    a. When asking questions with numeric answers or categorical questions with unknown universes of possible answers, ask open-ended questions rather than closed-ended questions.
    b. When life forces respondents to make choices outside of the survey context, study those choices via ranking questions rather than by asking people to rate individual objects.
4. To ensure consistent interpretation of rating scales, label all options with words and not numbers, and ensure that the range of options covers all points on the underlying continuum.
5. Break bipolar rating scales into multiple questions via branching, and offer unipolar rating scales without branching.
6. To optimally measure bipolar constructs, use seven-point response scales.
7. To optimally measure unipolar constructs, use five-point response scales.
8. Do not offer “don’t know” response options, and encourage respondents who volunteer a “don’t know” response to instead offer substantive answers.
9. Rotate the order of response options on rating scales and categorical questions across respondents, except when doing so would violate conversational conventions about order

(e.g., always put positive response options before negative options, order unipolar scales from most to least).

10. When selecting words, choose ones that are in frequent use in popular discourse, that have few letters and syllables, have only one primary definition (instead of two different frequently-used definitions), and are easy to pronounce.
11. Avoid asking people to remember the opinions they held at prior times – answers to such questions are usually wrong (unless the research goal is to assess what people believe their opinions used to be).
12. Avoid asking people to explain why they thought or behaved in particular ways – answers to such questions are usually wrong (unless the research goal is to assess why people think they thought or behaved in particular ways).

Perception of Visual Displays and Survey Navigation

1. In visual displays, avoid abbreviations, notation, and jargon (Kosslyn, 2007).
2. In visual displays, present no more than four visual elements (Kosslyn, 2007).
3. When presenting sequences of visual displays, identify pieces of new information with distinctive colors or sizes (Kosslyn, 2007).
4. Visual emphasis on certain words should be used sparingly, and bold typefaces and colors should be used instead of using all upper case words, italic typefaces, or underlining to create emphasis (Kosslyn, 2007).
5. Rely on the perceptual grouping laws of similarity and proximity to organize information presentation by grouping similar or related items together (Kosslyn, 2007).
6. Make important elements different from surrounding elements by making the former larger, brighter, or more distinctively colored (Kosslyn, 2007).

Pretesting Questionnaires

1. Implement "cognitive pretesting" of all questions, and change question wordings to eliminate misinterpretations or eliminate respondent confusion or lack of clarity of meaning (Willis, 2005).
2. Conduct multiple iterative rounds of cognitive pretesting to be sure that questionnaire revision does not introduce more error (Willis, 2006).
3. Behavior coding of interview administration can identify problematic questions that require revision.

Interviewer Deviations from the Standardized Survey Interviewing Script

1. To minimize interviewer deviations from scripted behavior, extensive training and monitoring should be implemented (Schaeffer et al., 2013).
2. To identify problematic behavior by interviewers, monitoring should include audio recordings and verbatim transcription (Schaeffer et al., 2013).

Coding Open-Ended Survey Questions

1. Multiple coders should code the same text independently, based on detailed written instructions, and the level of agreement between coders should be examined to confirm sufficient reliability. If reliability is insufficient, instructions may need to be revised, or coder training or supervision may need to be enhanced.
2. All materials used in coding should be made publicly available, including raw responses.

Confidentiality and Anonymity of Survey Participation and Data

1. Surveys should generally promise anonymity or confidentiality to respondents, which may produce higher participation rates and to minimize intentional misreporting.

Respondent Attrition from Panel Surveys

1. Repeated interviewing of the same respondents affords valuable analytic opportunities for studying change and causal processes (Schoeni et al., 2012).
2. Substantial effort should be expended to minimize panel attrition (Olsen, 2005).
3. Respondents who fail to provide data at one wave of a panel may rejoin the panel subsequently, so efforts to re-recruit such respondents are worthwhile.
4. Substantial paid incentives can often minimize panel attrition (Creighton et al., 2007).
5. Using information about individual respondents, tailor their incentive offers to be maximally attractive while not being unnecessarily high (Olsen, 2005).

Paradata – Data About Survey Processes and Contexts

1. Collect as much information about the process and context of the survey data collection as possible (Kreuter, 2013).
2. To evaluate operational aspects of web survey design features, collect timings for every item and page, which can be useful for analyzing cognitive processes (Heerwegh, 2003).
3. Use call record data for interviewer administered surveys to study non-response bias (Durrant et al., 2011).

Using Interviewer Observations About Respondents

1. To collect information about non-respondents, have interviewers record observable characteristics (age, gender, race, etc.) (Peytchev and Olson, 2007).
2. Have interviewers report evaluations of respondent engagement, perceptions of respondent honesty, perceptions of respondent ability (West, 2012).

Leave-behind Measurement Supplements

1. In-person interviewers can leave paper questionnaires with respondents, to be returned to investigators by mail, to collect supplementary measures (Harvey, 2002).

Capturing Moment-by-Moment Data

1. Experience sampling and ecological momentary assessment (EMA) allow for the collection of data in real time from respondents to supplement survey data (Stone and Shiffman, 1994).
2. To maximize accuracy of data collected using experience sampling or EMA, use very short recall periods (Stone et al., 1999).
3. Time-use diaries can be used to collect information on activities in real time (Bolger et al., 2003).

Collecting Biological Data via Biomarker Measures

1. To collect measures of health, use biomarkers (Schonlau et al., 2010).
2. To minimize costs of biomarker collection, dried blood spots (DBS) can be collected (McDade et al., 1999; Parker and Cubitt, 1999).

3. To collect extremely rich data, DNA can be collected, though analysis costs are high (Schonlau et al., 2010).

Specialized Tools for Measuring Past Events

1. To improve respondent recall of past events, especially with very long reference periods, use an event history calendar method to structure the interview (Belli, 1998).
2. To maximize the accuracy of respondent recall when using event history calendars, interviewers should be trained in cognitive retrieval strategies such as sequential retrieval, parallel retrieval, and top-down retrieval (Belli et al., 2007).

Linking Survey Data with Government Records

1. Linking survey data to official records on the same individuals has promise, but matching processes may not be as effective as is needed.
2. Materials used in matching should be made public to analysts, so the effectiveness of the matching can be evaluated.
3. If survey reports differ from measurements of the same phenomena in official records, consider the possibility that the records are incorrect rather than the self-reports (Presser et al., 1990).

Preserving and Maximizing the Accessibility of Existing Data

1. After a survey is conducted, the raw data and all instruments used to collect the data (e.g., the questionnaire, show cards, interviewer training manuals) should be preserved in electronic form and made available to the community of scholars (Ruggles et al., 2003).
2. Older surveys that exist on paper records only should be scanned and made electronically available in machine-actionable structured formats.
3. When related surveys have been conducted over time, the data from such surveys should be archived in a common, integrated format to facilitate comparison between surveys (Ruggles et al., 2003).
4. To move toward a sustainable model of data curation, survey data producers need to expand the use of persistent identifiers such as a Digital Object Identifier (DOI)[1].
5. Archiving of survey data and materials should be done using a common method, such as that prescribed by the Data Documentation Initiative (DDI)[2].

Improving Survey Website Usability

1. Websites that provide survey data and documentation to users should be designed based on the results of usability research to assure that users can easily find the information they require (U.S. Dept. of Health and Human Services, 2006).
2. Standardization of the format and content of such websites across survey projects would be desirable.

Research Transparency and the Credibility of Survey-Based Social Science

1. To maintain the credibility of survey data and survey-based empirical claims, researchers should be as transparent about research practices as possible (Lupia, 2008).

[1] http://www.doi.org/
[2] http://www.ddialliance.org/

2. To enable data users to evaluate assumptions made during data collection, processing, and analyses, survey data producers should clearly and uniformly document and make public all aspects of data production, including:
    - Sample selection
    - Respondent recruitment
    - Question selection
    - Question wording
    - Response options
    - Directions to interviewers
    - Post-interview processing (including construction of weights)
    - Coding of open-ended responses (Lupia, 2008)
3. AAPOR recommended standards for disclosure should be followed[3].

[3] http://www.aapor.org/Disclosure_Standards1.htm#.U2kEOq1dXzQ

Section 2 – Recommendations

In this section, we propose recommendations that NSF could implement to promote the standardization of the collection and distribution of data from the Foundation's major surveys in order to maximize data quality and consistency, as well as to reduce costs by avoiding duplication of activities that can most efficiently be done by a single entity.

1. In order to assure that data collection and dissemination procedures of SBE's major survey projects maximize quality, we recommend that SBE change the organization of its funding and monitoring of the projects, as follows:

    1a. We endorse the Committee of Visitors' recommendation that all major SBE survey projects be funded not from the disciplinary programs in which they have been housed and instead be funded through a new program managed by a single program officer.

    1b. We recommend that the new program officer appoint a data quality advisory panel to review and monitor the data collection procedures used by the major SBE survey projects.

    1c. We recommend that the new program exploit economies of scale, by avoiding duplication of activities across the surveys that can most efficiently be done by a single entity (e.g., have data dissemination experts set up the websites to document and disseminate data from the major surveys, rather than having each survey disseminate its own data based on idiosyncratic designs).

    1d. We recommend that the new program require that the teams of PIs who run each major SBE survey project include at least one individual whose research has been primarily in the methodology of collecting survey data and who has broad expertise in that arena.

    1e. We recommend that the new program require each SBE major survey project have a Board of Overseers that is funded via a separate grant to the chair of the board, not included in the cost of the main grant to the study PIs. The Board should report directly to SBE and provide independent review of the study's design and purposes on behalf of the interdisciplinary community of scholars who will use the data. Board members – some of whom should have expertise in survey methodology – should be selected by SBE.

    1f. We recommend that SBE review the recommendations offered by advisory and site visit committees created by SBE during the last 20 years to evaluate the methodologies of its surveys, to determine which recommendations of those committees have been followed, which have not, and whether further changes in procedures are warranted.

2. In order to provide better guidance on how to improve the quality of NSF surveys (as well as

surveys done by others), we recommend that SBE increase the funding of the MMS program so it can support substantially more research addressing challenges to survey methodology, including, for example:

- optimal computation of survey weights
- optimal procedures for pretesting survey questionnaires
- optimal procedures for training and supervising interviewers
- optimizing procedures for designing questions to minimize bias and maximize accuracy
- optimizing the use of official records data to complement survey data
- optimizing the use of incentives
- optimizing the coding of open-ended data
- optimizing procedures for validating that interviews were conducted properly

CONCLUDING REMARKS – Ways Forward for NSF

We appreciate the invitation from NSF to carry out this examination of contemporary survey methodology. Survey methods clearly continue to be of great value, and at the same time, there are many opportunities for improvement and enhancement of the method. We look forward to seeing NSF SBE leadership in this area to facilitate the development and dissemination of insights in this arena to improve the quality of understanding of modern societies.

REFERENCES:

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M.P., Couper, M.P., Dever, J.A., et al. (2013). Summary Report of the AAPOR Task Force on Non-Probability Sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143.

Belli, R.F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. *Memory*, *6*(4), 383-406.

Belli, R.F., Smith, L.M., Andreski, P.M., and Agrawal, S. (2007). Methodological Comparisons between CATI Event History Calendar and Standardized Conventional Questionnaire Instruments. *Public Opinion Quarterly*, *71*(4), 603–622.

Boehm, L.M. (1989). Reliability of Proxy Response in the Current Population Survey. In: *Proceedings of the Survey Research Methods Section*, (pp. 486-489). Alexandria, VA: American Statistical Association.

Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary Methods: Capturing Life as it is Lived. *Annual Review of Psychology*, *54*(1), 579–616.

Brick, J.M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, *29*(3), 359-362.

Brick, J.M., Brick, P.D., Dipko, S., Presser, S., Tucker, C., and Yuan, Y. (2007). Cell Phone Survey Feasibility in the U.S.: Sampling and Calling Cell Numbers Versus Landline Numbers. *Public Opinion Quarterly*, *71*(1), 23–39.

Church, A.H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly*, *57*(1), 62-79.

Cobb, C. and Krosnick, J.A. (2009). Experimental Test of the Accuracy of Proxy Reports Compared to Target Report with Third-Party Validity. Presented at the American Association for Public Opinion Research Annual Meeting, Hollywood, Florida.

Conrad, F.G., Broome, J.S., Benki, J.R., Kreuter, F., Groves, R.M., Vannette, D.L., and McClain, C. (2012). Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 191–210.

Creighton, K.P., King, K.E., and Martin, E.A. (2007). The Use of Monetary Incentives in Census Bureau Longitudinal Surveys. *Survey Methodology*, research report series no. 2007-2. Washington DC: U.S. Census Bureau.

de Leeuw, E.D. (2005). To Mix or not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, *21*(2): 233–255.

Dillman, D.A. and Christian, I.M. (2005). Survey Mode as a Source of Instability in Responses across Surveys. *Field Methods, 17*(1), 30–52.

Durrant, G.B., D'Arrigo, J., and Steele, F. (2011). Using Field Process Data to Predict Best Times of Contact Conditioning on Household and Interviewer Influences. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 174*(4), 1029–1049.

Durrant, G.B., Groves, R.M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly, 74*(1), 1–36.

Grice, H.P. (1975). Logic and Conversation. In: Davidson, D. and Harman, G. (Eds.), *The Logic of Grammar*, (pp. 64-75). Encino, CA: Dickenson Publishing Co.

Groves, R.M. and Couper, M.P. (2012). *Nonresponse in Household Interview Surveys*. New York, NY: John Wiley & Sons, Inc.

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 169*(3), 439–457.

Groves, R.M. and Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly, 72*(2), 167–189.

Harvey, A.S. (2002). Guidelines for Time Use Data Collection and Analysis. In: Pentland, W.E., Harvey, A.S., Lawton, M.P., and McColl, M.A. (Eds.), *Time Use Research in the Social Sciences*, (pp. 19-45). New York, NY: Kluwer Academic Publishers Group.

Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using Client-Side Paradata from a Web Survey. *Social Science Computer Review, 21*(3), 360–373.

Hox, J.J. and de Leeuw, E.D. (1994). A Comparison of Nonresponse in Mail, Telephone, and Face-To-Face Surveys. *Quality & Quantity, 28*, 329-344.

Keeter, S., Miller, C., Kohut, A., Groves, R.M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey. *Public Opinion Quarterly, 64*(2), 125–148.

Kosslyn, S.M. (2007). *Clear and to the Point: 8 Psychological Principles for Compelling PowerPoint Presentations*. New York, NY: Oxford University Press.

Kreuter, F. (Ed.). (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information*. New York, NY: John Wiley & Sons, Inc.

Krosnick, J.A. (1999). Survey Research. *Annual Review of Psychology, 50*(1), 537–567.

Link, M.W. and Mokdad, A.H. (2005). Advance Letters as a Means of Improving Respondent Cooperation in Random Digit Dial Studies: A Multistate Experiment. *Public Opinion Quarterly, 69*(4), 572–587.

Lupia, A. (2008). Procedural Transparency and the Credibility of Election Surveys. *Electoral Studies*, *27*, 732–739.

McDade, T.W., Williams, S., and Snodgrass, J.J. (2007). What a Drop Can Do: Dried Blood Spots as a Minimally Invasive Method for Integrating Biomarkers into Population-Based Research. *Demography, 44*(4), 899–925.

Olsen, R.J. (2005). The Problem of Respondent Attrition: Survey Methodology Is Key. *Monthly Labor Review*, *128*(2), 63-70.

Parker, S.P. and Cubitt, W.D. (1999). The Use of the Dried Blood Spot Sample in Epidemiological Studies. *Journal of Clinical Pathology*, 52(9), 633-*639.*

Peytchev, A. and Olson, K. (2007). Using Interviewer Observations to Improve Nonresponse Adjustments: NES 2004. In: *Proceedings of the Survey Research Methods Section,* (pp. 3364-3371). Alexandria, VA: American Statistical Association.

Presser, S., Traugott, M.W., and Traugott, S. (1990). Vote "Over" Reporting In Surveys: The Records or The Respondents. *American National Election Studies Technical Report Series*, No. nes010157. Available at http://www.umich.edu/~nes/.

Ruggles, S., King, M.L., Levison, D., and McCaa, R., and Sobek, M. (2003). IPUMS-International. *Historical Methods, 36*(2), 9–20.

Schaeffer, N.C., Garbarski, D., Freese, J., and Maynard, D.W. (2013). An Interactional Model of the Call for Survey Participation: Actions and Reactions in the Survey Recruitment Call. *Public Opinion Quarterly*, *77*(1), 323–351.

Schoeni, R.F., Stafford, F., McGonagle, K.A., and Andreski, P. (2012). Response Rates in National Panel Surveys. *The Annals of the American Academy of Political and Social Science*, *645*(1), 60–87.

Schonlau, M., Reuter, M., Schupp, J., Montag, C., Weber, B., Dohmen, T., et al. (2010). Collecting Genetic Samples in Population Wide (Panel) Surveys: Feasibility, Nonresponse and Selectivity. *Survey Research Methods, 4*(2), 121–126.

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in Household Surveys. In: Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (Eds.), *Survey Nonresponse*, (pp. 163-177). New York, NY: John Wiley & Sons, Inc.

Singer, E., Groves, R.M., and Corning, A.D. (1999). Differential Incentives: Beliefs about Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly, 63*(2), 251–260.

Stone, A.A. and Shiffman, S.S. (1994). Ecological Momentary Assessment (EMA) in Behavioral Medicine. *Annals of Behavioral Medicine, 16*, 199-202.

Stone, A.A., Shiffman, S.S., and DeVries, M.W. (1999). Ecological Momentary Assessment. In: Kahneman, D., Diener, E., and Schwarz, N. (Eds.), *Well Being: The Foundations of Hedonic Psychology*, (pp. 26–39). New York, NY: Russell Sage.

U.S. Dept. of Health and Human Services (2006). *The Research-Based Web Design and Usability Guidelines, Enlarged/Expanded edition*. Washington, DC: U.S. Government Printing Office. Available at: http://usability.gov/guidelines/guidelines_book.pdf.

Vicente, P. and Reis, E. (2010). Using Questionnaire Design to Fight Nonresponse Bias in Web Surveys. *Social Science Computer Review*, *28*(2), 251–267

West, B.T. (2012). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 211–225.

Willis, G.B. (2005). *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks, CA: Sage Publications.

Willis, G.B. (2006). Review: Cognitive Interviewing as a Tool for Improving the Informed Consent Process. *Journal of Empirical Research on Human Research Ethics, 1*(1), 9–24.

APPENDIX:

DOCUMENTATION OF CONFERENCES

Background

Survey research is at a crossroads. The need for information to track the public's behaviors, experiences, needs, and preferences has risen dramatically in recent years. Government agencies, businesses, academics, and others make decisions and create policies based on knowledge of populations, and a great deal of such information is collected via surveys. The unemployment rate, the inflation rate, and many other national indicators used widely in America are generated in this way. Thus, the need for high quality survey research is great and rising.

At the same time, the challenges of conducting high quality surveys are substantial. The U.S. Federal government remains committed to implementing face-to-face interviewing for many of its most important surveys, and in many countries around the world, face-to-face interviewing is the only way to reach a probability sample of the population without incurring substantial non-coverage. Furthermore, research to date suggests that face-to-face interviewing may be the method most likely to generate the highest response rates, the greatest trust and rapport between the researchers/interviewers and the respondents, the most cognitive effort from respondents in generating answers accurately, and the most honesty when providing reports regarding sensitive topics. But face-to-face interviewing is extremely expensive, and the costs of implementing such efforts well have been rising quickly.

In that light, alternative methods of data collection for surveys are appealing. Although telephone interviewing rose in popularity greatly in the 1970s as a more practical alternative to face-to-face interviewing, this method's response rates have been dropping in recent years, and costs have been rising. Remarkably, the accuracy of Random Digit Dial (RDD) telephone surveys appears to remain high, but respondents are less likely to be cognitively effortful and honest when being interviewed over the phone than when being interviewed face-to-face.

The rising costs of telephone interviewing set the stage for Internet data collection to become popular. And it has become so. Indeed, billions of dollars are spent annually around the world collecting survey data via the Internet. And some comparison studies have suggested that answering questions via a computer enhances cognitive performance and honesty relative to oral interviewing by telephone. When done with probability samples, Internet surveys seem to be a very promising avenue for effective and efficient data collection.

However, although a number of countries in addition to the U.S. now have commercial firms or academic institutions collecting survey data from probability samples of the population via the Internet (e.g., the Netherlands, Germany, France, Iceland, Norway, Sweden), this methodology has yet to catch on broadly across the world. Instead, most Internet survey data collection is done

from nonprobability samples of people who volunteer to complete surveys for money. Alternative vehicles, such as Amazon's Mechanical Turk, allow for such data collection from individuals who have not signed up to join a survey panel, and Google's survey platform allows for survey data collection from people who surf to a newspaper website and wish to continue reading a news story at no cost to them. Studies in the U.S. and abroad suggest that such data collection does not yield samples that reflect the population as accurately as do standard probability sampling methods.

But non-probability sample surveys implemented via the Internet have had tremendous appeal to researchers inside and outside of academia because of their practicality, especially their affordability. Thus, modes of data collection are in flux and in a state of tension. On the one hand, traditional, reliable methods are becoming increasingly costly. And on the other hand, new methods have obvious limitations in terms of their potential to produce generalizable results. At the same time, researchers are increasingly aware of another challenge in survey research: questionnaire design. For nearly a century, survey researchers have, for the most part, designed questionnaires in an intuition-driven, *ad hoc* fashion. As a result, there is tremendous heterogeneity in the design of questions across surveys and even within a single survey. Consider, for example, the ubiquitous rating scale, which has been used in countless surveys. The design of rating scales has no standardization across surveys – scales differ in terms of the number of points offered, the number of points that have verbal labels vs. numeric labels vs. no labels, the particular labels chosen, and the order in which the points are presented to respondents.

From this heterogeneity, an outside observer might conclude that there is no optimal way to design rating scales or, indeed, to make any other decisions when designing questionnaires. Instead, perhaps all question designs work equally well – as long as respondents can understand a question, they can answer it accurately, one might imagine.

But 70 years of research across the social sciences suggest that this is not true. In fact, hundreds if not thousands of studies provide guidance on how to design questions to maximize measurement reliability and validity, how to maximize the uniformity of respondent interpretations of questions, and how to minimize the cognitive demands made of respondents during the process of interpreting questions and answering them. But this information has yet to be disseminated and put into practice consistently across the nation's most important continuing and new surveys. Yet as practitioners' awareness of these best practices grows, so does concern about the value of data collected by questionnaires not conforming to these principles of optimizing measurement accuracy.

Furthermore, as the cost of survey data collection rises, other forms of data are increasingly available in the form of official records that some observers perceive to be potential replacements for survey data. That is, observers ask, "Why ask people whether they were victims of a crime when researchers can consult the electronic records of police departments to assess crime rates?" or "Why ask people how much they paid for milk when researchers can consult scanner data collected and retained by supermarkets?" The answers to these questions are actually quite simple in many cases: as appealing as these uses of official records are, those records are inadequate for many applications where survey data can serve the purpose effectively. For example, many crimes are not reported to the police, and some crimes reported to police

officers are not recorded in official records. So efforts to explore the full frequency of crimes require reports from people who experience them. Likewise, although supermarkets track purchases of products and can even link some purchases to the households that made the purchases, many purchases of food items are not made in such settings, and it is not yet possible to link purchasing behavior by a single individual across the full wide array of purchase settings without asking the individual via surveys. Thus, official records do not yet appear to be a viable replacement for all survey data.

Official records do appear to offer potential value in a different way: as a supplement to survey data. Consider, for example, the measurement of voter turnout. Agencies in almost all states in the country make available to researchers official records of who voted in each election, and some states provide a little additional information about the individuals, such as their gender and age. Furthermore, commercial companies offer services whereby they provide additional, in depth information purportedly about each individual on such lists, which can also be gathered from other publicly available records. These sorts of records are thought to be matchable to data collected from survey respondents to enrich understanding of these individuals with what are presumed to be accurate official records about them.

This accuracy hinges on the accuracy of the process by which a survey respondent is matched to official records purportedly about the same individual. This process of matching is being implemented by a number of commercial firms, but these firms consider the matching processes to be proprietary, so scientists cannot full observe the process and assess the accuracy of the results. It is possible that this matching process can be accomplished effectively using highly confidential federal government data obtainable via Census Data Centers, because individuals can be matched using their social security numbers. Advances in computer algorithms and computing power make this type of sophisticated and resource-intensive research increasingly achievable. However, very little research has actually examined these opportunities. Thus, the notion of enriching survey data with data from official records is both appealing and increasingly possible.

Another growing challenge in the survey research arena is the maintenance of records documenting how survey data were collected. In recent years, survey professionals have become increasingly sensitive to the importance of documenting absolutely all details of the process by which data collection occurs, to allow researchers to understand the data and differences in results obtained by different data collection methods. This includes show cards displayed to respondents, interviewer training manuals, text of open-ended questions, detailed field reports to permit calculation of response rates using various contemporary methods, and much more information collected by survey researchers and often not retained or disseminated in ways that allow for in-depth, accurate understanding by scholars. Recently, the dissemination of survey data and survey data collection documentation has advanced considerably. But most survey research organizations are not collecting and disseminating information about their surveys optimally. As a result, analysts are handicapped, uninformed about important aspects of the process by which data were generated (and therefore unable to tailor analysis accordingly) and unable to explore important design issues that might impact findings.

All of the above issues and more were explored in the two NSF-sponsored conferences described

in this report. Experts outlined the challenges faced by survey researchers, opportunities to strengthen the methodology, and immediate steps that can be taken to enhance the value of the methodology to a wide range of users, in academia, government, and the private sector.

Presenters

Kathleen T. Ashenfelter - U.S. Census Bureau

Kathleen T. Ashenfelter is the principal researcher and group leader for the Human Factors and Usability Research Group in the Center for Survey Measurement at the U.S. Census Bureau. Dr. Ashenfelter earned a Ph.D. in Quantitative Psychology from the University of Notre Dame in 2007. Her research interests include human-computer interaction, analysis of human interaction, eye-tracking methodology, and survey rostering research.

Robert Belli - University of Nebraska-Lincoln

Robert Belli is Director of the Survey Research and Methodology Program and UNL Gallup Research Center and Professor in the Department of Psychology at the University of Nebraska-Lincoln. He served as North American Editor of Applied Cognitive Psychology from 2004-2009. He received his Ph.D. in experimental psychology from the University of New Hampshire in 1987.

Dr. Belli's research interests focus on the role of memory in applied settings, and his published work includes research on autobiographical memory, eyewitness memory, and the role of memory processes in survey response. The content of this work focuses on false memories and methodologies that can improve memory accuracy. Current research is examining the electrophysiological correlates of suggestibility phenomena. Teaching interests include courses on basic and applied cognitive psychology, and on the psychology of survey response.

Matthew Berent – Matt Berent Consulting

Since receiving his Ph.D. in social psychology in 1995 from The Ohio State University, Matthew Berent has held a variety of academic and private sector jobs across the United States. The diversity of his work history has helped him develop a broad range of skills that includes basic and applied research, teaching students and professionals, product development and management, and sales and marketing.

Dr. Berent's professional interests include exploring the factors that lead to more reliable and valid survey data, developing a methodology for measuring emotional reactions in real time during human-computer interactions, and refining methods for capturing voice of customer data.

Bob Blaemire – Catalist

Bob Blaemire is Director of Business Development at Catalist, a voter database organization. Mr. Blaemire has been an active participant in politics his entire adult life. Born and raised in Indiana, his career began at the age of 18 upon entering George Washington University. His employment with Senator Birch Bayh (D-IN) began in 1967 during Bob's freshman year and concluded with

Bayh's unsuccessful re-election campaign in 1980. Those 13 years saw Mr. Blaemire rise from volunteer worker to Office Manager to Executive Assistant in the Senate Office, and, finally, Political Director of the 1980 campaign.

After the 1980 defeat, he founded a political action committee, The Committee for American Principles, an organization seeking to combat the growing role and influence of the New Right in political campaigns. He began his career providing political computer services in 1982, eventually joining with and starting the Washington Office of Below, Tobe & Associates. In 1991, Mr. Blaemire created Blaemire Communications, a political computer services firm serving Democratic campaigns, progressive organizations and political consultants. During that time, Blaemire Communications managed more Democratic state party voter file projects than any other vendor. In late 2007, Blaemire Communications was acquired by Catalist.

J. Michael Brick - Westat

Michael Brick is a Vice President of Westat, where he is co-director of the Survey Methods Unit and the Statistical Staff. He is also a research professor in the Joint Program in Survey Methodology at the University of Maryland. His Ph.D. is in Statistics from American University.

Dr. Brick has served as: President, Washington Statistical Society (2010-2011); co-chair of AAPOR Task Force on Non-Probability Sampling (2011-); Statistics Canada Statistical Methodology Committee (2007- present); Editorial Board for Public Opinion Quarterly (2003-present); Publication Editor for International Association of Survey Statisticians Jubilee Commemorative Volume, Landmark Papers in Survey Statistics (2001); and, Editor for The Survey Statistician (1995-2000). Dr. Brick is a fellow of the American Statistical Association and an elected member of the International Statistical Institute. His professional interests include: survey sampling and methodology, nonresponse bias, and modes of data collection. He has published numerous articles in various survey and statistical journals on these topics.

Curtiss Cobb – Facebook

Curtiss Cobb is manager of the Growth Population and Survey Sciences Group at Facebook. Previously, he was Senior Research Director and Director of Survey Methodology at GfK Custom Research North America (formerly Knowledge Networks). He received a Ph.D. in Sociology from Stanford University in 2010 and was a research fellow at the Stanford Institute for the Quantitative Study of Society and a graduate research fellow at the Center for Survey Measurement at the U.S. Census Bureau.

Dr. Cobb's fields of research include organizational behavior, science and technology, professions, survey design and methodology, public opinion and political behavior. Dr. Cobb also received a M.A. in Quantitative Methods for Social Sciences from Columbia University and a B.A. in Political Science and Psychology from the University of Southern California.

Matthew DeBell - Stanford University

Matthew DeBell has been director of Stanford Operations for the American National Election Studies since 2006. He received his Ph.D. in Government from Georgetown University. Before coming to Stanford University, he worked for the American Institutes for Research in Washington, D.C., where he

conducted survey research and data analysis for the National Center for Education Statistics. Dr. DeBell's research areas include questionnaire design, survey weighting, and public opinion.

Robert Groves - Georgetown University

Robert Groves is provost of Georgetown University. Dr. Groves is a social statistician who studies the impact of social cognitive and behavioral influences on the quality of statistical information. His research has focused on the impact of mode of data collection on responses in sample surveys, the social and political influences on survey participation, the use of adaptive research designs to improve the cost and error properties of statistics, and public concerns about privacy affecting attitudes toward statistical agencies.

Prior to joining Georgetown as provost he was Director of the U.S. Census Bureau (Presidential appointment with Senate confirmation), a position he assumed after being director of the University of Michigan Survey Research Center, Professor of Sociology, and Research Professor at the Joint Program in Survey Methodology at the University of Maryland. Dr. Groves is an elected fellow of the American Statistical Association, elected member of the International Statistical Institute, elected member of the American Academy of Arts and Sciences, elected member of the US National Academy of Sciences, and elected member of the Institute of Medicine of the US National Academies.

Scott Keeter - Pew Research Center

Scott Keeter is Director of Survey Research for the Pew Research Center in Washington DC. He is co-author of four books, including A New Engagement? Political Participation, Civic Life, and the Changing American Citizen (Oxford University Press), The Diminishing Divide: Religion's Changing Role in American Politics, (Brookings Institution Press), What Americans Know about Politics and Why It Matters (Yale University Press), and Uninformed Choice: The Failure of the New Presidential Nominating System (Praeger). His other published research includes articles and book chapters on survey methodology, political communication and behavior, and health care topics.

Dr. Keeter is past president of the American Association for Public Opinion Research, and previously served as chair of the Standards Committee and Councilor-at-Large for the Association. Since 1980 he has been an election night analyst of exit polls for NBC News.From 1998 to 2002 he was chair of the Department of Public and International Affairs at George Mason University, and previously taught at Rutgers University and Virginia Commonwealth University, where he also directed the Survey Research Laboratory from 1988-1991. A native of North Carolina, he attended Davidson College as an undergraduate and received a Ph.D. in political science from the University of North Carolina at Chapel Hill.

Frauke Kreuter - University of Maryland

Frauke Kreuter is a professor in the Joint Program in Survey Methodology (JPSM) at the University of Maryland. She received her Master in Sociology from the University of Mannheim, Germany and her Ph.D. in Survey Methodology from the University of Konstanz. Before joining the University of Maryland, she held a postdoc at the UCLA Statistics Department. Her research focuses on sampling and measurement errors in complex surveys. In her work at JPSM she maintains strong ties to the Federal Statistical System, and served in advisor roles for the National Center for Educational Statistics and the Bureau of Labor

Statistics.

Dr. Kreuter has authored five books and edited three others, along with publishing numerous journal articles and reports. She is a fellow of the American Statistical Association, Standards Chair for the American Association for Public Opinion Research, and the Co-chair for the Task Force on Big Data for the American Association for Public Opinion Research.

Jon Krosnick - Stanford University

Jon Krosnick is the Frederic O. Glover Professor in Humanities and Social Sciences and professor of communication, political science, and psychology at Stanford University. Winner of the American Association for Public Opinion Research's Lifetime Achievement Award, he is a leading international authority on questionnaire design and survey research methods, and on the psychology of attitudes, especially in the area of politics. For 30 years, over the course of seven books and more than 190 articles and chapters, Dr. Krosnick has studied how the American public's political attitudes are formed, change, and shape thinking and action.

His research explores the causes of people decisions about whether to vote, for whom to vote, whether to approve of the President's performance, whether to take action to influence government policy-making on a specific issue, and much more. His recent research has focused on how other aspects of survey methodology (e.g., collecting data by interviewing face-to-face vs. by telephone or on paper questionnaires) can be optimized to maximize accuracy. He has also been conducting national surveys on global warming for more than 15 years and has monitored and interpreted changes in public perception over time on this issue.

Dr. Krosnick is co-principal investigator of the American National Election Study, the nation's preeminent academic research project exploring voter decision-making and political campaign effects. Dr. Krosnick's scholarship has been recognized with the Phillip Brickman Memorial Prize, the Pi Sigma Alpha Award, the Erik Erikson Early Career Award for Excellence and Creativity, two fellowships at the Center for Advanced Study in the Behavioral Sciences, and membership as a fellow of the American Academy of Arts and Sciences and the American Association for he Advancement of Science.

Gary Langer – Langer Research Associates

The founder and president of Langer Research Associates, Gary Langer is an internationally recognized public opinion researcher with expertise in analysis of political, policy, economic and social attitudes, questionnaire design, data interpretation, survey methodology and survey management. With more than 25 years in the field, Mr. Langer has overseen and analyzed more than 700 surveys on a broad range of topics as director of polling at ABC News. Mr. Langer has won two Emmy awards and received nine Emmy nominations – including the first and only to cite public opinion polls – as well as being honored with the 2010 Policy Impact Award of the American Association for Public Opinion Research for a series of surveys in Afghanistan and Iraq, described in AAPOR's citation as "a stellar example of high-impact public opinion polling at its finest." He's a two-time winner of the University of Iowa-Gallup Award for Excellent Journalism Using Polls, produced a pair of ABC News polls recognized by the Excellence in Media Coverage of Polls Award from the National Council on Public Polls and shared a DuPont-Columbia Award for ABC's 9/11 coverage.

Mr. Langer created ABC's industry-leading survey standards and vetting operation and has advanced disclosure initiatives through professional organizations. A frequent speaker, writer and commentator on public attitudes, he is the author of an award-winning blog, the "The Numbers," at ABCNews.com, has authored or co-authored nearly 30 scholarly papers and has given scores of invited presentations on the meaning and measurement of public opinion. Mr. Langer is a member of the Board of Directors of the Roper Center for Public Opinion Research, a trustee of the National Council on Public Polls and past president of the New York chapter of the American Association for Public Opinion Research. A Phi Beta Kappa graduate of the University of New Hampshire, he lives in New York with his wife and two daughters.

Mark Liberman - University of Pennsylvania

Mark Liberman is the Christopher H. Browne Distinguished Professor of Linguistics at the University of Pennsylvania, where he is also a professor of Computer and Information Science and a member of the Psychology Graduate Group. From 1975 to 1990, he worked as a member of technical staff and head of the Linguistics Research Department at AT&T Bell Laboratories. In 1992, he founded the Linguistic Data Consortium (LDC), which creates, collects and distributes speech and text databases, lexicons, and other resources for language-related education, research and technology development. He is a frequent contributor to "Language Log", a popular linguistics weblog.

Michael W. Link - The Nielsen Company

Michael Link is Chief Methodologist for Research Methods at The Nielsen Company. He has a broad base of experience in survey research, having worked in academia (University of South Carolina, 1989-1999), not-for-profit research (RTI International, 1999-2004), government (Centers for Disease Control and Prevention, 2004-2007), and the private sector (Nielsen, 2007-present). He received a Ph.D. in Political Science from the University of South Carolina.

Dr. Link's research centers around developing methodologies for confronting some of the most pressing issues facing survey research, including techniques for improving survey participation and data quality (use of address-based sampling, impact of call screening technologies), methodological issues involving use of multiple modes in data collection (web, mail, CATI, field, mobile, meters), and obtaining participation from hard-to-survey populations (linguistically isolated, racial and ethnic groups). His numerous research articles have appeared in Public Opinion Quarterly and other leading scientific journals. Dr. Link is the President of the American Association for Public Opinion Research

Arthur Lupia - University of Michigan

Arthur Lupia is the Hal R. Varian Collegiate Professor of Political Science and a research professor at the Institute for Social Research at the University of Michigan. Dr. Lupia examines how information and institutions affect policy and politics, with a focus on how people make decisions when they lack information. He draws from multiple scientific and philosophical disciplines and uses multiple research methods. His topics of expertise include information processing, persuasion, strategic communication, and civic competence. He has published widely. He has held a range of scientific leadership positions including Principal Investigator of the American National Election Studies. Dr. Lupia also has developed new means for researchers to better serve science and society. As a founder of TESS

(Time-Sharing Experiments in the Social Sciences; www.experimentcentral.org), he has helped hundreds of scientists from many disciplines run innovative experiments on opinion formation and change using nationally representative subject pools.

He is regularly asked to advise scientific organizations and research groups on how to effectively communicate science in politicized contexts. He currently serves an executive member of the Board of Directors of Climate Central. He is past president of the Midwest Political Science Association and past chair of the Association for the Advancement of Science's Section on Social, Economic and Political Sciences. His many articles appear in political science, economics, and law journals, and his editorials are published in leading newspapers. His research has been supported by a wide range of groups including the World Bank, the Public Policy Institute of California, the Markle Foundation, and the National Science Foundation. He has also received multiple honors including the National Academy of Sciences' Award for Initiatives in Research.
He is an elected member of the American Academy of Arts and Sciences, a fellow of the American Association for the Advancement of Science. He has been a Guggenheim Fellow and a recipient of the National Academy of Science's Award for Initiatives in Research.

Michael P. McDonald – University of Florida

Michael McDonald is an associate professor of Political Science at the University of Florida and a Non-Resident Senior Fellow at the Brookings Institution. He received his Ph.D. in Political Science from University of California, San Diego and B.S. in Economics from California Institute of Technology. He held a one-year post-doc fellowship at Harvard University and has previously taught at Vanderbilt University and University of Illinois, Springfield.

Dr. McDonald's research interests include voting behavior, redistricting, Congress, American political development, and political methodology. His voter turnout research shows that turnout is not declining, the ineligible population is rising. He is a co-principle investigator on the Public Mapping Project, a project to encourage public participation in redistricting. He is co-author with Micah Altman and Jeff Gill of *Numerical Issues in Statistical Computing for the Social Scientist* and is co-editor with John Samples of *The Marketplace of Democracy: Electoral Competition and American Politics*. His research appears in several edited volumes and in scholarly journals.

Dr. McDonald has worked for the national exit poll organization, consulted to the U.S. Election Assistance Commission, consulted to the Pew Center for the States, served on campaign staff for state legislative campaigns in California and Virginia, has worked for national polling firms, has been an expert witness for election lawsuits in Florida and Washington, and has worked as a redistricting consultant or expert witness in Alaska, Arizona, California, Georgia, Michigan, New Jersey, New York, Ohio, Oklahoma, Texas, and Virginia. He has worked as a media consultant to ABC and NBC, and is frequently quoted in the media regarding United States elections. His opinion editorials have appeared in *The Washington Post*, *The Politico, The Milwaukee Journal-Sentinel*, *The American Prospect*, and *Roll Call*.

Colm O'Muircheartaigh - University of Chicago

Colm O'Muircheartaigh, is a professor at the University of Chicago Harris School of Public Policy. He served as dean of the Harris School from 2009 to 2014. His research encompasses survey sample design, measurement errors in surveys, cognitive aspects of question wording, and latent variable models for nonresponse. He is a senior fellow in the National Opinion Research Center

(NORC), where he is responsible for the development of methodological innovations in sample design.

Dr. O'Muircheartaigh is co-principal investigator on NSF's Center for Advancing Research and Communication in Science, Technology, Engineering, and Mathematics (ARC-STEM) and on the National Institute on Aging's National Social Life Health and Aging Project (NSHAP). He is a member of the Committee on National Statistics of the National Academies (CNSTAT) and of the Federal Economic Statistics Advisory Committee (FESAC), and serves on the board of Chapin Hall Center for Children.

Dr. O'Muircheartaigh joined Chicago Harris from the London School of Economics and Political Science (LSE), where he was the first director of the Methodology Institute, the center for research and training in social science methodology, and a faculty member of the Department of Statistics from 1971. He has also taught at a number of other institutions, having served as a visiting professor at the Universities of Padova, Perugia, Firenze, and Bologna, and, since 1975, has taught at the Summer Institute of the University of Michigan's Institute for Social Research.

Formerly president of the International Association of Survey Statisticians and a council member of the International Statistical Institute, Dr. O'Muircheartaigh is actively involved in these and a number of other professional bodies. He is a fellow of the Royal Statistical Society, a fellow of the American Statistical Association, and an elected member of the International Statistical Institute. He was a member of the U.S. Census Bureau Federal Advisory Committee of Professional Associations (chair of the statistics subcommittee), a member of the Advisory Boards of the Panel Study on Income Dynamics (PSID) and the National Longitudinal Study of Adolescent Health (Add Health), and a member of the National Academies Panel on Residence Rules for the 2010 Census. He has served as a consultant to a wide range of public and commercial organizations in the United States, the United Kingdom, Ireland, Italy, and the Netherlands. Through his work with the United Nations (FAO, UNDP, UNESCO), OECD, the Commission of the European Communities, the International Association for Educational Assessment (IEA), and others, Dr. O'Muircheartaigh has also worked in China, Myan Mar, Kenya, Lesotho, and Peru.

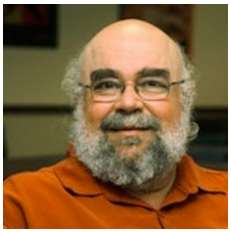
Randall Olsen - Ohio State University

Randall Olsen is a professor of Economics and the Director of the Center for Human Resource Research (CHRR) and has been Principal Investigator with the National Longitudinal Surveys of Labor Market Experience (NLS) for over twenty years. He also works on other projects at CHRR, most of which are related to applied microeconomics. For example, CHRR has operated the Consumer Finance Monthly survey since 2005, which has collected detailed data on household balance sheets, financial management and financial knowledge. Having been continuously collected for over two years in advance of the financial panic and subsequent recession as well as since then, these data provide a unique lens through which researchers can study the recent financial dislocations. The Consumer Finance Monthly is the only continuously operated household financial survey of American households over this period.

Dr. Olsen also has a project investigating depressivity over the life cycle and across generations using NLS data collected over the past several decades, including the child development study funded by NIH – the Children of the NLSY79. Another current project looks at the hedonic valuation of traits in the marriage market, allowing for the possibility that men and women place differential values and traits. This project also aims to infer whether trait valuations are different for different cohorts, comparing people born 1957-64 with those born 1980-84.

Dr. Olsen maintains long-standing interests in econometrics, labor economics, applied micro and economic demography. He has served a term as the Director of the Initiative in Population Research – an NICHD funded interdisciplinary population center at Ohio State. He guided that effort for four of its first five years and helped establish it as a thriving center for social science research that spans a great many colleges and departments at Ohio State.

Steven Ruggles - University of Minnesota



Steve Ruggles, Regents Professor of History and Population studies, Director of the Minnesota Population Center, and Distinguished McKnight University Professor in the College of Liberal Arts, has been at the University of Minnesota since 1985. He is also President of the Population Association of America.

Dr. Ruggles is a "scholar of astonishing breadth of knowledge and productivity whose work has reshaped the field of historical demography and has had a profound effect on the fields of sociology, economics, and history." He is considered to be one the most widely known historical demographers in the world. He is credited with raising approximately $65 million in research funds to create, improve, and disseminate population data. He and his team have collected, coded, computerized, systematized, and delivered to scholars the individual records of millions of Americans going back to 1850. This high-precision individual-level census database for the United States is known as the Integrated Public Use Microdata Series (IPUMS).

More recently, he has formed international collaborations to expand IPUMS to a global scale. IPUMS is the single most widely used data source in the top-ranked journal of population research. Nothing like it has ever existed. It may be the most valuable database of all time and it has made the University of Minnesota one of the largest social science data distributors in the world! He provided key leadership in two other large-scale collaborative data infrastructure projects - the North Atlantic Population Project (NAPP), which extended the chronological reach of international data collections; and the National Historical Geographic Information System (NHGIS), which provided web-based access to all U.S. Census summary data since 1790 and integrated electronic boundary files describing the historical locations of counties and census tracts. These data collections are believed to have had a profound and a transformational effect on social science research.

Dr. Ruggles has also produced pioneering studies of historical change in demographic behavior. His work on the history of the American family is described as "path-breaking." He is the author of numerous publications on demographic and family change and censuses and data. His monograph received the William J. Goode Distinguished Book Award by the American Sociological Association and the Allen Sharlin Memorial Award by the Social Science History Association. He is known to be an extraordinary teacher who has had a profound impact on undergraduate and graduate education. He has served as adviser to twelve Ph.D. students and has worked with fifteen postdoctoral research associates. His service to the University and professional organizations is outstanding. He has served on more than thirty-five University-wide, collegiate, and department committees and over seventeen professional boards and committees. In addition, he has served on editorial boards and served as a peer reviewer of grant proposals.

Joe Sakshaug - University of Mannheim

Joseph Sakshaug is an assistant professor of Statistics and Social Science Methodology in the Department of Sociology at the University of Mannheim, senior researcher in the Department of Statistical Methods at the Institute for Employment Research (IAB) in Nuremberg, and a faculty associate in Survey Methodology at the University of Michigan. He received M.S. and Ph.D. degrees from the Program in Survey Methodology at the University of Michigan in 2007

and 2011, respectively. His research interests include administrative data linkage, statistical disclosure control, applications of multiple imputation, and nonresponse and measurement error in surveys.

Nora Cate Schaeffer - University of Wisconsin, Madison

Nora Cate Schaeffer is Sewell Bascom Professor of Sociology at the University of Wisconsin, Madison, where she also serves as faculty director of the University of Wisconsin Survey Center, teaches courses in survey research methods, and conducts research on questionnaire design and interaction during survey interviews.

Dr. Schaeffer serves as member of the Public Opinion Quarterly Advisory Board of the American Association for Public Opinion Research and of the General Social Survey Board of Overseers. She recently (2009) completed terms as the Council on Sections Representatives for the Survey Research Methods Section of the American Statistical Association and as a member of the Census Advisory Committee of Professional Associations. Her service for the National Research Council (NRC) includes the Panel on the Design of the 2010 Census Program of Evaluations and Experiments, the Committee on National Statistics, the Panel to Review Research and Development Statistics at the National Science Foundation, and the Panel to Evaluate Alternative Census Methods for the National Research Council. Other previous service includes the American Statistical Association Technical Advisory Committee on the Survey of Income and Program Participation; the Technical Review Committee for the National Longitudinal Survey of Youth; the National Science Foundation Advisory Committee for the Social, Behavioral, and Economic Sciences; and the governing Council of the American Association for Public Opinion Research. Dr. Schaeffer has also served on the editorial boards of Public Opinion Quarterly, Sociological Methods and Research, and Sociological Methodology. Before receiving a Ph.D. from the University of Chicago (1984), she worked at National Opinion Research Center. In 2010, she was selected as a fellow of the American Statistical Association.

Eleanor Singer - University of Michigan

Eleanor Singer has been an academic survey researcher at the Survey Research Center at the Institute for Social Research at the University of Michigan, where she is Research Professor Emerita. Beginning in 1975, the focus of her work has been the empirical investigation of ethical issues in the treatment of human subjects. Starting with a study of the consequences of informed consent procedures for response rates and response quality in surveys, to which she has returned at various points in her career, she has studied the impact of concerns about privacy and confidentiality on participation in a variety of settings and in face-to-face, telephone, mail, and Web surveys. Dr. Singer has also examined the effect of incentives on response rates, expectations about future rewards, and behavior in future surveys, as well as the potentially “coercive” effects of using monetary rewards to induce survey participation. All of these studies have looked at actual effects--of requesting signed consent, providing information to respondents about sensitive content, giving greater or lesser assurances of confidentiality, and offering greater or lesser rewards, or none at all—in order to better inform the ethical and effective practice of survey research. In addition to her own research, she has served on or chaired various National Academy panels on privacy and confidentiality, data access, and protection of human research subjects.

After graduating from Queens College with a degree in English in 1951, Dr. Singer worked as an editor at various publishing houses, including Teachers College Press, and increasingly specialized in books about social science—a fact that prompted her return to graduate school

at Columbia University in 1959. There, contact with Lazarsfeld and Merton, and in particular with her dissertation sponsor Herbert H. Hyman, introduced her to public opinion research and survey methodology, AAPOR, and Public Opinion Quarterly.

Before becoming active in AAPOR, Dr. Singer edited Public Opinion Quarterly from 1975-86, and was largely responsible for gaining AAPOR's sponsorship of POQ as its official journal, a move that assured the journal's financial stability and editorial continuity. She was President of AAPOR from 1987-8; Conference Chair from 1984-5, Standards Chair from 1995-6, and Counselor-at-large from 1985-7 and 1991-3. She received the AAPOR Award for Lifetime Achievement in 1996.

Arthur Stone – University of Southern California

Arthur Stone is a professor Psychology and director of the Center for Self-Report Science at the University of Southern California. Dr. Stone specializes in the field of behavioral medicine, and he has conducted studies on stress, coping, physical illness, psychoneuroimmunology, psychoendocrinology, structured emotional writing, and self-report processes. Many of his studies have used diaries and momentary approaches to data capture. Dr. Stone's current research focuses on the properties of momentary data in the context of pain and chronic illnesses. In his role as a Gallup Senior Scientist, which began in 2005, Dr. Stone is working with Gallup researchers to explore how employee engagement relates to workers' physical health and wellbeing.

Dr. Stone's coauthored book titles include *The Science of Self Report* and *The Science of Real-Time Data Capture*. His most recent journal contributions include: "Understanding Recall of Weekly Pain From a Momentary Assessment Perspective: Absolute Agreement, Between- and Within-Person Consistency, and Judged Change in Weekly Pain," "A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method (DRM)," and "Variability of Momentary Pain Predicts Recall of Weekly Pain: A Consequence of the Peak (or Salience) Memory Heuristic." Dr. Stone has been an executive council member for the American Psychosomatic Society; a research committee member for the American Psychological Association; and a past president and executive council member of the Academy of Behavioral Medicine Research. His editorial appointments include editor-in-chief for *Health Psychology* and *Annals of Behavioral Medicine*; editorial board member for the *British Journal of Health Psychology* and *Mind/Body Medicine*; and journal reviewer for more than 15 psychology publications. A licensed psychologist, Dr. Stone received his bachelor's degree from Hamilton College and doctorate degree in clinical psychology from Stony Brook University. Recent honors and awards include the Distinguished Health Psychologist Senior Award from the American Psychological Association, Division 38; the University Medal of the University of Trier, Germany; and becoming a SUNY Distinguished Professor.

Roger Tourangeau – Westat

Roger Tourangeau is a vice president and associate director in the Statistics Group and co-director of Westat's Survey Methods Group. He is also editor of the Journal of Survey Statistics and Methodology from the American Association for Public Opinion Research. With some 30 years of experience as a survey methodologist, Dr. Tourangeau has designed and conducting studies on a wide range of topics and is well known for his research on the impact of different modes of data collection on the answers and on the cognitive processes underlying survey responses. He is the lead author of The Psychology of Survey Response, which received the 2006 AAPOR Book Award, and he was given the 2002 Helen Dinerman Award, the highest honor of the World Association for Public Opinion Research for his work on cognitive aspects of survey methodology.

He also has a recent book *The Science of Web Surveys* (with Frederick Conrad and Mick Couper). Dr. Tourangeau is an expert on questionnaire design in surveys and has taught courses on this subject. He is a research professor emeritus at the University of Michigan. Before coming to Westat, Dr. Tourangeau worked at NORC, the Gallup Organization, and the University of Michigan; during his stint at the University of Michigan, he served as the Director of the Joint Program in Survey Methodology for nine years. He has a Ph.D. in psychology from Yale University and is fellow of the American Statistical Association.

David Vannette - Stanford University

David Vannette is a Ph.D. Candidate in the Department of Communication at Stanford University. His general research interests are in the areas of survey methodology, political psychology, and political communication. A key question that he examines in his research is how people interpret and use data about public opinion to inform their own attitudes and behaviors towards subjects. Prior to beginning his Ph.D. research, David received a B.S. from Calvin College and an M.S. in Survey Methodology from the University of Michigan.

Mr. Vannette's research has been published in scholarly journals such as the *Journal of the Royal Statistical Society: Series A* and the *Joint Statistical Meetings Proceedings*, three book chapters, and four case studies at the University of Michigan. His experience includes work for the Institute for Social Research, Survey Research Center, Ross School of Business, and the William Davidson Institute at the University of Michigan and United States Census Bureau. He has worked as a teaching assistant for the Summer Institute in Political Psychology and other courses at Stanford in addition to assisting in teaching MBA courses in the Marketing and Strategy departments at the University of Michigan Ross School of Business. More recently, he has presented research advice and insights for private research firms, non-profit organizations, agencies of the U.S. Government, and federal government contractors.

David Weir - University of Michigan

David Weir is a research professor at the Survey Research Center and research affiliate with the Population Studies Center of the Institute for Social Research at the University of Michigan. Dr. Weir's current research interests include the measurement of health-related quality of life; the use of cost-effectiveness measures in health policy and medical decision-making; the role of supplemental health insurance in the Medicare population; the effects of health, gender, and marital status on economic well-being in retirement; and the effects of early-life experience on longevity and health at older ages. He also directs the Health and Retirement Study (HRS). He holds a Ph.D. in Economics from Stanford University and was previously an associate professor of Economics at Yale University.

Brady West - University of Michigan

Brady West is a research assistant professor in the Survey Methodology Program, located within the Survey Research Center at ISR. He also serves as a statistical consultant at the Center for Statistical Consultation and Research (CSCAR) on the University of Michigan-Ann Arbor campus. He earned his Ph.D. from the Michigan Program in Survey Methodology in 2011. Before that, he received an M.A. in Applied Statistics from the U-M Statistics

Department in 2002, being recognized as an Outstanding First-year Applied Masters student, and a B.S. in Statistics with Highest Honors and Highest Distinction from the U-M Statistics Department in 2001.

His current research interests include the implications of measurement error in auxiliary variables and survey paradata for survey estimation, survey nonresponse, interviewer variance, and multilevel regression models for clustered and longitudinal data. He is the lead author of a book comparing different statistical software packages in terms of their mixed-effects modeling procedures (*Linear Mixed Models: A Practical Guide using Statistical Software*, Chapman Hall/CRC Press, second edition, 2014), and he is a co-author of a second book entitled *Applied Survey Data Analysis* (with Steven Heeringa and Pat Berglund), which was published by Chapman Hall in April 2010.

Gordon Willis - National Cancer Institute/National Institutes of Health

Gordon Willis is a cognitive psychologist in the Office of the Associate Director of the Applied Research Program. Prior to that, Dr. Willis was Senior Research Methodologist at Research Triangle Institute, and he also worked for over a decade at the National Center for Health Statistics, CDC, to develop methods for developing and testing survey questions.

Dr. Willis attended Oberlin College, and received a Ph.D. in Cognitive Psychology from Northwestern University. He now works mainly in the area of the development and evaluation of surveys on cancer risk factors, and focuses on questionnaire pretesting. He has produced the "Questionnaire Appraisal System" for use in evaluating draft survey questions, and has written the book *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. He also co-teaches a graduate-level questionnaire design course at the Joint Program for Survey Methodology at the University of Maryland, and serves as Adjunct Faculty at the Uniformed Services University of the Health Sciences (USUHS). His research interests have recently turned to cross-cultural issues in self-report surveys and research studies, and in particular the development of best practices for questionnaire translation, and the development of pretesting techniques to evaluate the cross-cultural comparability of survey questions. Dr. Willis also works in the area of human subjects protection in cancer research, and has served as Chair of the NCI Special Studies Institutional Review Board (IRB). He is a member of the Editorial Board of the Journal of Empirical Research on Human Research Ethics, and consults regularly on matters pertaining to ethical issues in population-based research.

Summaries of Presentations

Section 1: Conventional Survey Research

Reasons for Optimism about the Accuracy of Survey Research – Jon A. Krosnick

Although research on the accuracy of surveys is valuable and exists in many corners of many disciplines, no concatenation of this knowledge exists. Many articles and books have focused on survey errors resulting from issues relating to coverage, sampling, nonresponse, and measurement, but very little work has comprehensively evaluated survey accuracy and can serve as a touch point for researchers and data users interested in answering a simple question: how accurate are survey measurements?

Assessing survey accuracy is challenging because such work requires having an external measure of the "true" values of a variable in order to be able to judge how well that value is measured by a survey. For example, in the area of voting behavior, self-reports of turnout are often collected in surveys and compared with the official turnout statistics provided by the Federal Election Commission (FEC) after the election. When these sources yielded different rates, the discrepancy has often been assumed to be attributable to mistakes in the self-reports; the FEC numbers are assumed to document the truth.

Chang et al. (2011) conducted a review of many studies that have assessed survey accuracy. This review constitutes the first-ever meta-analysis of survey accuracy. The authors identified four principal methods for assessing the accuracy of survey results and collected published studies using each method. These studies assessed accuracy in a wide range of domains, including behaviors in the arenas of healthcare utilization, crime, voting, media use, smoking, and more, and measures of respondent characteristics such as demographics, height, and weight.

The authors identified 555 studies that matched each respondent's self-report data with objective individual records of the same phenomena, resulting in a dataset of over 520,000 individual matches. This method of verification indicated that for over 85 percent of the measurements, there was perfect agreement between the survey data and the objective records or measures. Second, the investigators found 399 studies that matched one-time aggregate survey percentages and means with available benchmarks from non-survey data. These studies involved different units of measurement, such as percentages, means in centimeters, kilograms, days, hours, drinks, etc. This assessment method indicated that survey measures matched benchmarks exactly in 8 percent of the instances, 38 percent manifested almost perfect matches (less than 1 unit difference), and 73% manifested very close matches (less than 5 units difference). Third, the authors found 168 instances in which studies correlated individuals' self-reports on surveys with secondary objective data. The results obtained using this method indicated generally strong positive associations between the self-reports and the secondary data. The authors identified 6 studies that correlated trends over time in self-reports and with trends in objective benchmarks. This approach documented very strong associations between the self-report survey data and trends in the objective benchmarks. Thus, in this meta-analysis, Chang and her colleagues examined over 1000 published comparisons assessing the accuracy of survey data, and the

results indicated that the vast majority of survey measurements of objective phenomena were extremely accurate.

When differences do occur between survey estimates and objective benchmarks, it is important to consider why these differences may have arisen rather than immediately discounting the survey data. For example, some observers have assumed that surveys overestimate voter turnout because of respondent lying. That is, respondents are thought to believe that voting is socially desirable, and so people who didn't vote may claim to have voted in order to look presentable or to avoid revealing an embarrassing fact about themselves. However, the accumulating literature suggests instead that individual survey reports of turnout may be remarkably accurate, and the problem may be that people who participate in elections also over-participate in surveys. If so, the disagreement between aggregate rates of turnout according to surveys vs. government statistics may not be due to inaccurate respondent reporting and may instead be due to a variety of other factors, including problems in matching records to respondents.

These findings should give survey producers, consumers, and funding agencies considerable optimism about the continued accuracy of surveys as a method of collecting data. The findings also indicate that survey research deserves its role as one of the most used and trusted methods for data collection in the social sciences.

Reference:

Chang, L., Krosnick, J.A., and Albertson, E. (2011). How Accurate Are Survey Measurements of Objective Phenomena? Paper Presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix, AZ.

Probability vs. Non-probability Sampling Methods – Gary Langer

Before 1936, data on populations generally were collected either via a census of the entire population or "convenience" sampling, such as straw polls. The latter, while quick and inexpensive, lacked a scientific, theoretical basis that would justify generalization to a broader population. Using such methods, the Literary Digest correctly predicted presidential elections from 1916 to 1932, but the approach collapsed in 1936. The magazine sent postcards to 10 million individual selected from subscriptions, phone books and automobile registration records. Through sampling and self-selection bias, the 2.4 million responses disproportionately included Republicans, and the poll predicted an easy win for the losing candidate, Alf Landon.

George Gallup used quota sampling in the same election to draw a miniature of the target population in terms of demographics and partisanship. Using a much smaller sample, Gallup correctly predicted Franklin D. Roosevelt's win. This set the stage for systematic sampling methods to become standard in polling and survey research.

But quota sampling turned out not to be a panacea. The approach suffered a mortal blow in the 1948 presidential election, when Gallup and others erroneously predicted victory for Thomas Dewey over Harry Truman. While a variety of factors were responsible, close study clarified the shortcomings of quota sampling. Replicating the U.S. population in terms of cross-tabulations by

ethnicity, race, education, age, region, and income, using standard categories, would require 9,600 cells, indicating a need for enormous sample sizes. Further, "The microcosm idea will rarely work in a complicated social problem because we always have additional variables that may have important consequences for the outcome" (Gilbert et al., 1977). And bias can be introduced through interviewers' purposive selection of respondents within each quota group.

After spirited debate, survey researchers coalesced around probability sampling as a scientifically rigorous method for efficiently and cost-effectively drawing a representative sample of the population. In this technique, each individual has a known and ideally non-zero probability of selection, placing the method firmly within the theoretical framework of inferential statistics. As put by the sampling statistician Leslie Kish, "(1) Its measurability leads to objective statistical inference, in contrast to the subjective inference from judgment sampling, and (2) Like any scientific method, it permits cumulative improvement through the separation and objective appraisal of its sources of errors" (Kish, 1965).

In modern times, high-quality surveys continue to rely on probability sampling. But new non-probability methods have come forward, offering data collection via social media postings and most prominently though opt-in online samples. These often are accompanied by ill-disclosed sampling, data collection, and weighting techniques and yet also routine claims that they produce highly accurate data. Such claims need close scrutiny, on theoretical and empirical bases alike.

Opt-in surveys typically are conducted among individuals who sign up to click through questionnaires on the Internet in exchange for points redeemable for cash or gifts. Opportunities for falsification are rife, as is the risk of a cottage industry of professional survey respondents. One study found that among the 10 largest opt-in survey panels, 10 percent of panelists produced 81 percent of survey responses, and one percent of panelists accounted for 24 percent of responses.

The approach raises many questions. Who joins these poll-taking clubs, what are their characteristics, and what do we know about the reliability and validity of their responses? Are respondent identities verified? Are responses validated? What sorts of quality control measures are put in place? What survey weights are applied, how were they obtained, and what is their effect? What claims are made about the quality of these data, and how are these claims justified?

An example of further challenges in opt-in online surveys is their common and generally undisclosed use of routers to maximize efficiency of administration, albeit at the cost of coverage. As an illustration, participants may be asked if they are smokers; if so, are routed to a smoking survey. If not smokers, they may be asked next if they chew gum. If yes, they are routed to a gum-chewers survey. If not, they may next be asked if they use spearmint toothpaste, and so on. Unbeknownst to sponsors of the toothpaste study, smokers and gum chewers are systematically excluded from their sample.

Purveyors of opt-in online sampling often point to the declining response rates and increasing costs of probability-based telephone surveys, topics that are addressed later in this report. But these arguments are hardly a constructive defense of alternative methodologies, nor do they recognize the wealth of research identifying response rates as a poor indicator of data quality.

Rather than pointing toward potential deficiencies in existing methods, it seems incumbent on the proponents of alternative non-probability methods to construct a reasoned defense of the approach, including a theoretical basis for its validity.

Empirical research consistently has found validity in scientific probabilistic sampling methods. Results for non-probability opt-in panels have been more concerning. An extensive review of existing literature, the AAPOR Report on Online Panels, published by the American Association for Public Opinion Research in 2010 (Baker et al., 2010), recommended that "researchers should avoid nonprobability online panels when one of the research objectives is to accurately estimate population values." This report concluded, "There currently is no generally accepted theoretical basis from which to claim that survey results using samples from nonprobability online panels are projectable to the general population. Thus, claims of 'representativeness' should be avoided when using these sample sources." (Subsequent to this presentation, an AAPOR report on non-probability sampling, in 2013, again noted the absence of a theoretical framework that would support statistical inference (Baker et al., 2013)).

In a large empirical study in 2011, Yeager and his colleagues compared seven opt-in online sample surveys with two probability sample surveys, finding that the probability surveys were "consistently highly accurate" while the opt-in samples were "always less accurate… and less consistent in their level of accuracy" (Yeager et al., 2011). The authors also found little empirical support for the claim that some non-probability panels are consistently more accurate others. They reported that weighting did not always improve accuracy of these panels, and they found no indication that higher completion rates produce greater accuracy. A report on data produced for a study by the Advertising Research Foundation found similar problems, as did an independent analysis of 45 individual data-quality studies (Baker, 2009; Callegaro et al., 2012). These confirm the fundamental issue: the absence of theory that would predict accurate, reliable results from non-probability samples.

Even if they can't be used to generalize about a broader population, it has been suggested that non-probability approaches are sufficient for evaluating associations among variables and for tracking trends over time. However, an empirical study on propensity to complete the U.S. Census, comparing otherwise identical probability-based and non-probability surveys, indicated otherwise. It found "systematic and often sizable differences between probability sample telephone data and non-probability Internet data in terms of demographic representativeness of the samples, the proportion of respondents reporting various opinions and behaviors, the predictors of intent to complete the Census form and actual completion of the form, changes over time in responses, and relations between variables" (Pasek and Krosnick, 2010). More study is warranted, but the picture to date is bleak.

Another recent trend is to evaluate information made publicly available on social networks such as Facebook and Twitter. The appeal of these datasets is their size and scope. Data can be collected on a minute-by-minute basis in vast quantities on nearly any topic imaginable. While these forms of data may hold great potential for social scientists, they also present unique challenges. For example, it may be assumed that a Twitter or Facebook post represents one individual expressing his or her actual opinion on something once. In fact some users may post multiple times, using a single account or multiple accounts. Postings may not reflect the self-

initiated expression of actual attitudes, but rather may be part of orchestrated campaigns. Accounts may be created by interest groups, corporations, or paid public relations agents. Posts may be produced by automated computer programs known as "bots". Fake accounts can be purchased in bulk. All of these forms of information exist within the same datasets.

Regardless of their source, selecting relevant postings and extracting meaning from them are further challenges. Many postings include slang, irony, sarcasm, abbreviations, acronyms, and emoticons, or lack identifiable context. Tests of automated coding systems indicate highly inconsistent results. And again we face the lack of theoretical justification to make inferences about a broader population.

What does the future hold for non-probability samples? Can they be "fixed"? Some researchers suggest the use of Bayesian adjustment, or a return to sample matching. While further research is welcome, what has been lacking to date is the required transparency that must underlie any such evaluation. Non-probability methods should be held to the same analytical standards and evaluated on the same basis as probability samples with regard to claims of accuracy, validity, and reliability. Full disclosure of methods and data quality metrics is crucially important. And the Holy Grail remains the development of an online sampling frame with known probabilities of selection, bringing the enterprise into harmony with sampling theory.

Probability sampling requires ongoing evaluation as well. Some organizations implement poor-quality sampling designs and suboptimal execution and analysis. Coverage is an ongoing concern, and the potential impact of declining response rates needs continuing examination. So does work on probability-based alternatives to traditional telephone methods, such as address-based sampling, mixed-mode designs, and others that may be developed in the future.

Areas for future research:

- Expanded empirical research into the validity and reliability of non-probability survey data
- Efforts to develop a theoretical framework under which such samples may support inference
- Improved assessment and methods of analysis of social media data
- Continued examination of probability-based methods
- Development and implementation of transparency standards

References:

Baker, R. (2009). Finally, The Real Issue? [July 7 blog post]. The Survey Geek. Available at http://regbaker.typepad.com/regs_blog/2009/07/finally-the-real-issue.html right.

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., et al. (2010). Research Synthesis: AAPOR Report on OnLine Panels. *Public Opinion Quarterly, 74*(4), 711-781.

Baker, R., Brick, J.M., Bates, N.A., Battaglia, M.P., Couper, M.P., Dever, J.A., et al. (2013). Summary Report of the AAPOR Task Force on Non-Probability Sampling. *Journal of Survey Statistics and Methodology, 1*(2), 90–143.

Callegaro, M., Villar, A., Krosnick, J.A., and Yeager, D. (2012). A Systematic Review of Studies Investigating the Quality of Data Obtained with Online Panels. Paper presented at the 67th annual conference of the American Association of Public Opinion Research, Orlando, FL.

Gilbert, J.P., Light, L.R., and Mosteller, F. (1977). Assessing Social Innovations: An Empirical Base for Policy. In: Fairley, W.B. and Mosteller, F. *Statistics and Public Policy*. Reading, MA: Addison-Wesley Pub Co.

Kish, L. (1965). *Survey Sampling*. New York, NY: John Wiley & Sons, Inc.

Pasek, J. and Krosnick, J.A. (2010). Measuring Intent to Participate and Participation in the 2010 Census and Their Correlates and Trends: Comparisons of RDD Telephone and Non-Probability Sample Internet Survey Data. *Survey Methodology Report* #2010-15, Washington, DC: Statistical Research Division, U.S. Census Bureau.

Yeager, D., Krosnick, J.A., Chang, L.C., Javitz, H.S., Levendusky, M.S., Simpser, A., and Wang, R. (2011). Comparing The Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. *Public Opinion Quarterly*, *75*, 709-747.

Sampling for Single and Multi-Mode Surveys using Address-based Sampling - Colm O'Muircheartaigh

A combination of factors in recent years provides an opportunity to transform the application of survey sampling methodology in the U.S. The potential is particularly exciting in multi-mode surveys. Survey sampling at its most basic level involves identifying and selecting potential sample members from a population using a sampling frame; the sampling frame comprises the set of materials – lists, maps, etc. – that best covers the population that we wish to describe or analyze, the population of inference.

The perfect sampling frame should include each element in the population once, and only once, and include only eligible population elements. A sampling frame can suffer from two primary types of deficiencies: overcoverage or undercoverage. Undercoverage has traditionally been the greater problem; here the sampling frame fails to include eligible population elements. Overcoverage occurs when the frame includes elements that appear to be members of the population but turn out not to be. Consider the case of a survey being conducted by telephone. A frame of listed landline telephone numbers would fail to include unlisted landline numbers and cell-phone numbers and thus exclude those whose telephone access was from an unlisted number or by cellphone only. If we chose instead to include all possible 10-digit numbers in the frame (or even all seven-digit numbers within existing area codes), the great majority of the numbers in the frame would have no working telephone associated with them. These redundant numbers would consist of frame overcoverage.

Consider the general problem of frame construction for a survey of the household-residing U.S. population. We want to develop a set of materials that will enable us first to select a sample of residents (or households) and then recruit them to participate in a survey. [There are three basic survey modes that we use: mail (self-completion) and two interviewing modes, telephone and face-to-face.] Every element of the household-residing population can be associated with the location (and address) of their household. If we could construct a frame containing all these addresses, we would have the foundation of a very robust sampling frame. Recent developments have made this a real (though still imperfect) prospect.

Nearly all households in the United States have a mailing address that is used by United States Postal Service (USPS) to deliver mail. This full set of addresses (technically a very close facsimile) can be obtained in the form of the USPS computerized Delivery Sequence File (DSF or CDSF). The DSF has been designed as an organizational tool for the USPS and enables them to effectively route the delivery of mail at every subdivision of the organization down to the individual mail carriers. The DSF is continuously updated by individual mail carriers via "edit books" that allow them to keep their routes updated with any changes such as new or deleted addresses. As it is in the interest of the mail carrier to have an accurate list of addresses, we have considerable confidence in the general quality of the carrier route data; our methodological research suggests that this belief is well founded. The entire DSF is updated on a monthly basis, and can be obtained through third-party vendors who have licensing arrangements with the USPS.

The DSF has nearly 100% coverage of households in the U.S. The DSF does contain addresses that are not easily identified or located; some examples are Post Office boxes, rural route boxes, drop points, and vacant dwellings; though the incidence of such cases is decreasing, they still present non-trivial challenges.

As a potential sampling frame, the DSF has excellent coverage properties and is greatly superior to the alternatives. The DSF can be used as a frame for many modes, given the proper conditions. For mail it is extremely efficient and effective; it is designed for the mail system. For face-to-face interviewing, it is also excellent and is actually superior in many ways to traditional sample frame that needed to be built and tested using listing methods. For telephone sampling, it is effective only when the address sampled can be readily matched to a telephone number for that sample unit (which may be an organization, household, or individual). As the basis for a web sample, the DSF frame has limited current value, as it is considerably more challenging to match e-mail addresses to mailing addresses than even telephone numbers. However, in the future more sophisticated approaches to matching may enable ABS to be used for web surveys.

Given the versatility of ABS for different modes, it is a particularly powerful approach when multi-mode methods are appropriate. Multi-mode surveys are increasingly common and important in survey research given coverage and nonresponse issues with telephone surveys and the escalating costs of face-to-face interviewing. The renaissance in mail surveys has also been a factor. Multi-mode approaches are becoming more common in an attempt to overcome, or bypass, the problems associated with single-mode methods.

In order to be able to conduct a multi-mode survey, the samples of respondents need to be drawn from identical or at least comparable frames and the sample design needs to be compatible in order to be able to transfer and control cases. These requirements make the DSF a likely frame for multi-mode studies, particularly when telephone numbers can be matched to the DSF. Furthermore, advances in technology have made compatibility across modes more feasible and computers are now more able to handle the complex decision rules and branching involved. However, multi-mode approaches are not without their own important drawbacks including susceptibility to different mode effects and an increase in the complexity of data collection and analyses.

While the DSF has presented a great opportunity to develop ABS over the past ten years, there are still a number of limitations that need to be addressed by future research. Over- and undercoverage are the main limitations; while the DSF is better than any other potential frames on these dimensions, there is room for improvement. With regard to overcoverage, identifying units that no longer exist, businesses that have been misclassified as residential, and misgeocoded addresses are problems that will be examined by future research to see if there are solutions that survey researchers can implement. Undercoverage is a larger problem due to the previously mentioned issue of drop points; additional issues are rural vacancies, new construction, simplified addresses that lack the standard formatting needed for sampling, and incomplete address fields such as ZIP codes.

Not all problems affect the DSF uniformly. An important example is the issue of “drop points”, places where the mail carrier drops off the mail and then someone else distributes it to the individual residences. Drop points are most common for multi-unit residences such as apartment complexes. Drop points represent only about two percent of the total addresses on the DSF frame, but in the cities where these drop points tend to be clustered, they present a significant problem. In Chicago, approximately fifteen percent of addresses are at drop points, and in New York City, the percentage is close to twenty percent. As the DSF does not specify names associated with an address, it is currently impossible to conduct mail surveys at drop points, and as the mail identifier is only at the drop point level, it becomes almost impossible to match any other data to individual households residing at that drop point, making the matching of telephone numbers to the particular household impossible also unless supplementary data are available. For face-to-face surveys, the problem is remediable as field interviewers can carry out a listing and select a household at the drop point.

Future research should focus on ways to augment the DSF in ways that may be useful to survey researchers. For example, vendors such as InfoUSA and Experian can add demographic and other information at the address level. This can include household composition, race/ethnicity, and income, along with market-derived data such as car ownership, magazine subscriptions, and other purchasing behavior. With the caveat that such supplementary information may be subject to serious measurement errors, any information that can be obtained outside the survey could enrich both the design (through stratification for instance) and the analysis (by providing ancillary variables for control or explanation). Identifying novel approaches to augmenting the DSF promises to be an extremely useful and fruitful area of future research.

Research on reconciling frames in order to make them more compatible will allow ABS to become even more useful as the basis for multi-mode approaches. Linking frames will enable better coverage and make sampling easier, making the entire survey process more flexible and versatile. It is also important to develop hierarchical and relational data structures within which it will be easier to switch modes and survey instruments, even within units, dynamically during data collection without invalidating the analyses. Linking frames with sophisticated database management approaches will enable rapid responses to changes when a survey is in the field. Building in capacity to use an adaptive approach to partially completed cases or requests within cases to change modes could help boost response rates and reduce expense. In the multi-mode context, where ABS is perhaps most useful, it is important to continue work on questionnaire design and the development of comparable stimuli for different modes; this will include work on mode effects of question form and question order sequences and their implications for the reliability and validity of data collected across modes.

Areas for future research:

- Novel ways of exploiting the value of ABS for a broader range of surveys
- Augmenting the DSF with ancillary data
- Addressing the problem of drop points on the sampling frame
- Identifying improved methods for reconciling and linking disparate sampling frames

The Impact of Survey Nonresponse on Survey Accuracy – Scott Keeter

Over the past 25 years, there has been a consistent and significant decline in response rates to surveys. This can be tracked most reliably in long-term surveys that have decades of year-over-year data on the response rates that they have achieved. For example, the National Household Education Survey has gone from a response rate of 60 percent down to nearly 30 percent in the eleven-year period from 1996 to 2007. In the 1970s and 1980s, conventional wisdom suggested that the quality of inference from survey data would decline rapidly once response rates dropped below 50%. Yet today response rates above 60 percent are the exception rather than the rule, even for the most important national surveys such as the "Big 3" funded by the NSF: the American National Election Studies (ANES), General Social Survey (GSS), and Panel Study of Income Dynamics (PSID).

Fortunately, survey results have maintained a remarkable level of reliability despite declining response rates. The consistent high level of performance by high quality surveys in the face of this challenge is a testament to the robustness of the survey research paradigm. This robustness has done little to quell the perennial fears associated with declining response rates in the survey research community. The great fear is that at some unknown point response rates will degrade to the level where no amount of post hoc adjustment will reduce nonresponse bias enough to use the data for population inference.

However, a very compelling meta-analysis by Groves and Peytcheva (2008) demonstrated that at any given level of nonresponse the level of bias varies considerably, meaning that nonresponse itself does not reliably predict nonresponse bias. Subsequent experimental research has supported

this notion by demonstrating that higher response rates achieved by increasing expense and effort in reducing nonresponse via refusal conversion did not provide different estimates than the same surveys conducted at lower effort/cost/response rates.

While this is good news for the validity of survey data in the face of declining response rates, it does raise another problem: Within any given survey at any level of nonresponse, there can be significant variability in the amount of nonresponse bias for individual measures. This presents a serious concern because it means that researchers have to figure out what kinds of measures have the greatest likelihood of being biased by nonresponse. This is further complicated by the fact that the nonresponse bias can be caused by a number of different factors, including survey design features and characteristics of respondents, both of which may interact to create another layer of causal complexity to disentangle. This makes predicting nonresponse bias incredibly difficult, and indeed there is no comprehensive theory of survey response that can generate reliable predictions about when nonresponse bias will occur. Because nonresponse bias is so unpredictable, it has also been very difficult to generate remedies or a set of best practices aimed at preventing it from occurring.

There are a few areas for future research that the NSF can support that will help develop a better understanding of the impact that declining response rates have on the accuracy of survey data. First, survey researchers should take a concerted look at smarter ways to invest in reducing nonresponse bias. Rather than focusing on propping up unsustainable and largely irrelevant nominal response rates, we should be asking what promising areas of survey design or opportunities for data integration and matching across databases might provide a better return on investment? Second, more basic research is needed into the correlates of nonresponse bias both in terms of survey design and respondent characteristics. One promising but under-utilized design in this regard is seeding samples with households that have known characteristics and then observing their response propensities to provide more precise estimates of nonresponse bias. Finally, future research should examine the promise of developing novel weighting schemes based on different known characteristics of respondents and nonrespondents. For example, volunteering behavior has been associated with response propensity and could be used as an important variable when creating post-stratification weights. If more variables like this can be identified, then nonresponse bias in a greater variety of outcome variables can be estimated more precisely and corrected for more adequately.

Areas for future research:

- Identifying targeted approaches to combating non-response during survey administration (e.g., adaptive design) that may yield better insurance against non-response bias than simply applying comparable effort toward all non-responding cases.
- Identifying novel weighting approaches based on known non-demographic characteristics of respondents and nonrespondents

Reference:

Groves, R. M. and Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. *Public Opinion Quarterly*, *72*(2), 167–189.

Optimizing Response Rates – J. Michael Brick

As discussed in the section above, response rates are a major concern for all researchers conducting surveys. Considerable work has been done to seek ways to stem the tide of declining response rates and also to examine the impact of response rates on data quality. The federal government often requests that funded projects obtain 'optimal response rates' in Requests for Proposals (RFPs), but the very notion of an optimal response rate is ill defined. There are two broad conceptualizations that we might infer from this wording: (1) simply maximize the response rate, or (2) achieve the response rate that minimizes nonresponse bias.

In the first conceptualization, researchers are interested in identifying survey design features that will maximize overall response rates within a fixed data collection cost. Key factors in this regard are the content and salience, the sponsor of the survey, and the mode in which the survey is conducted. The content and sponsor of the survey may be relevant to some respondents, influencing the salience of the survey and their willingness to participate. Mode is often one of the most important decisions that influences both cost and response rates. However, these factors are usually challenging to alter. Salience issues are often idiosyncratic to particular respondents, and it is cost that determines the mode more often than the desired response rate.

Beyond these largely fixed factors, there are some design features that may aid researchers in maximizing response rates. Refusal conversion is particularly important regardless of survey mode; refusal conversion refers to the process of attempting to convert sampled individuals who have previously refused to participate in the survey to become respondents. The number of contacts made to a sampled household is another key to increasing the response rate. Staff training, whether for interviewers or other survey staff, is also important. Interviewer training has been demonstrated to increase response rates. For mail and web surveys, appropriate design of the questionnaires and other ancillary survey materials such as letters and invitations can aid in maximizing response rates. The relative importance of each of the factors outlined above may vary depending on the type of survey and the population being sampled. For example, households may differ on many of the design features and dimensions that influence response rates when compared with organizational surveys or surveys of other specialized populations such as teachers or doctors.

For cross-sectional household surveys, there are a few best practices for achieving the biggest increase in response per dollar spent. First, a token monetary incentive can substantially increase response rates. The demands of the survey and burden on the respondent should be considered when determining the amount of the incentive; for example, a survey collecting medical specimens such as blood or saliva should provide larger incentives, as should surveys that require a significant amount of the respondent's time to complete.

The design of the survey materials is often overlooked but can have a significant impact on survey response rates. Poorly designed surveys may increase breakoffs, as may surveys that require respondents to perform a very uninteresting task at the outset such as an extensive household roster. Other best practices include developing plans for training interviewers, the number and protocol for contact attempts, and refusal conversion. A protocol that includes multiple contact attempts is critical to obtaining higher response rates. Advance letters

informing the respondent of their impending invitation to participate in the survey have also been demonstrated to improve response rates.

To maximize overall response rates, it is important to realize that choices of these influencing design factors are interrelated. The total survey design approach of Don Dillman and the leverage-salience theory of survey participation suggest that the factors may influence different respondents in varied ways. Thus, the point is not to define a set of rigid rules defining the survey recruitment process, but to recognize that many factors are at work and different combinations may be necessary to maximize overall response rates.

If the goal is to maximize overall response rates, then likely strategy involves "cherry picking" respondents, that is, targeting respondents with the highest propensity to respond. While the survey organization may not explicit have "cherry picking" is an objective, when the main message to interviewers is to increase response rates, the way the interviewers may respond is to choose the easiest households to reach this goal. This strategy results in obtaining more responses from people in the following demographic groups: older, female, homeowners, English-speakers, and upper-middle income. These demographics are typically associated with the highest willingness to participate for the lowest amount of effort. To increase response rates it is possible interviewers may target more of these people. However, this sort of approach is unlikely to reduce nonresponse bias, which should be the primary concern when considering response rates.

Often survey clients, including government organizations, will request a particular response rate without acknowledging that higher response rates may actually increase nonresponse bias. This type of fixed response rate requirement places survey providers in the position of recruiting more of the same types of people that typically respond to surveys. The survey provider must meet the requirement, even if this is likely to exacerbate nonresponse bias.

A more scientifically valid approach to response rates is to focus on achieving the response rates that minimize nonresponse bias. This is a more difficult construct and harder to incorporate in a fair way into the survey requirements. However, considerable evidence exists demonstrating that response rates do not reliably predict nonresponse bias. Thus, instead of simply trying to maximize response rates or specifying a particular arbitrary and high response rate, survey clients should shift their focus to minimizing nonresponse bias. This is not an easy task, and it is much harder to evaluate.

Early research on nonresponse bias modeled it as a deterministic function of the differences between respondent and nonrespondent characteristics and the nonresponse rate. However, this model depends on having measured nonrespondent characteristics, which may be expensive, impractical, and is not commonly possible. A more modern best practice involves modeling nonresponse bias as a stochastic process based on the association between response propensity and the characteristic being estimated. In this model, nonresponse bias cannot be viewed as simply a function of nonresponse rates. It is a much more nuanced and complex problem involving (1) response propensities of units, (2) the type of estimates being derived, (3) the specific estimator, and (4) the auxiliary data used as covariates to model the nonresponse bias.

Future research is needed to understand the reasons that nonresponse bias occurs and what the causal mechanisms are. There may be particular indicators of nonresponse bias that researchers should regularly measure and document. These could be features of the survey such as sponsorship, content, mode, questionnaire design, etc., or they could be characteristics of respondents, such as altruism.

Because surveys produce many types of estimates, and bias is a function of the particular statistic, future research to improve methods for optimizing nonresponse bias is not a simple task. A large body of research has looked at the impact of declining response rates on nonresponse error and described many features of nonresponse bias. One area that deserves more attention is to predict nonresponse bias quantitatively – specifically when bias will occur and its magnitude. Existing theory provides very little insight into the underlying measurable causes of nonresponse bias, and empirical research has been largely unable to produce high levels of bias even under conditions where it is expected. One possible explanation is that the unpredictability is related to the dependencies associated with design features affecting response. Thus, future research should aim to develop a more comprehensive theory of nonresponse bias that generates testable hypotheses and reliably predicts the conditions under which nonresponse bias will be observed in the real-world context.

Concrete steps for future research should include comparative analysis of respondents and nonrespondents, with a focus on features that can be experimentally manipulated. Nonresponse bias is likely to continue to be a significant and intractable problem until researchers are able to reliably produce it in an experimental context. Additionally, theories of survey response should be tested in practice; producing differential response rates that are in line with theory are a necessary place to start. It would be helpful if such research begins with low-cost factors such as the design of survey materials. On the analysis side, more research is needed that evaluates potential links between statistical adjustments that are performed post hoc and data collection procedures. It would be extremely beneficial to know if steps can be taken in the data collection process that reduce the need for statistical adjustments later on.

Progress toward understanding and reducing nonresponse bias is likely to remain limited without the development of a broad program of research aimed specifically at this area. Individual studies that are not linked by a central program of research are not likely to be sufficient. NSF should consider funding a cohesive set of projects or a larger program of research aimed at understanding nonresponse in a more comprehensive and systematic manner.

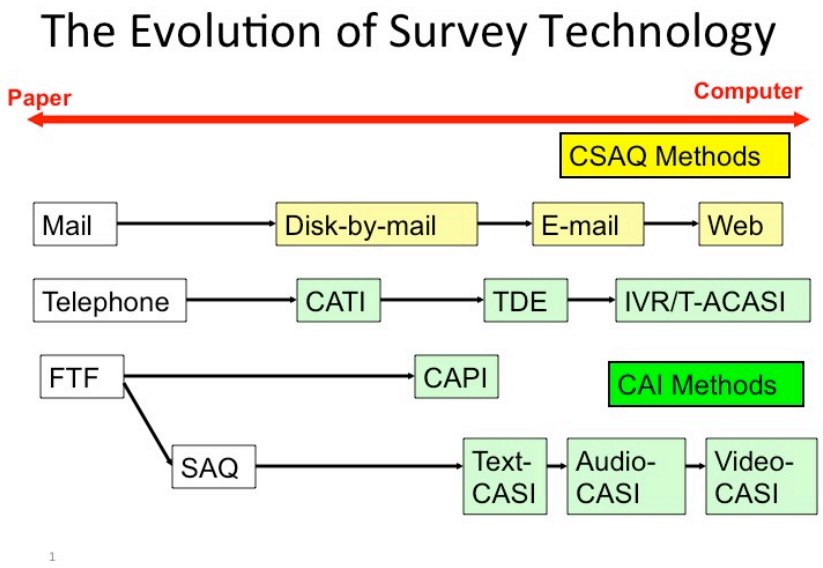
In summary, some areas for future research include:

- Developing a comprehensive program of research aimed at understanding and measuring nonresponse bias
- Comparative analysis of respondents and nonrespondents focused on features that can be experimentally manipulated
- Procedures to test nonresponse bias theories in practice:
    - Programs that aim to produce differential response rates for domains as predicted
    - Programs that aim to change domain response rates based on a sequence of actions
    - Focus on low-cost factors such as material design

- Evaluations linking statistical adjustments and data collection procedures

Modes of Data Collection – Roger Tourangeau

Survey research began in the late 1930s and the early 1940s, and for the first several decades of its history, virtually all surveys used just two modes of data collection: mail questionnaires or face-to-face interviews. As telephone coverage of the population improved over time, different sectors of the survey research industry began adopting telephone interviewing as a way to improve response rates relative to mail surveys and to decrease costs relative to face-to-face surveys. Different sectors of the survey research industry adopted telephone interviewing at different times, with the surveys done by or for the federal statistical system being the last to adopt the mode in the 1970s.



This chart, which was adapted from one done by Mick Couper, shows the evolution in data collection modes, beginning in the 1970s with the three traditional modes: mail, telephone, and face-to-face (FTF). Since then, there have been two major waves of change. In the first, computers began replacing paper as the basic medium on which surveys were conducted. For example, with the telephone surveys, by the mid-1970s parts of the survey research industry had already begun switching to computer-assisted telephone interviewing (CATI). This method became very popular, and CATI essentially came to replace paper telephone surveys almost entirely. Similarly, as computers got smaller and lighter, some survey organizations began sending interviewers out with computers for use with face-to-face surveys, an approach that came to be known as computer-assisted personal interviewing (CAPI), and it eventually supplanted paper face-to-face interviews almost completely as well. Surveys that had traditionally used mail were slower to adapt to computerization, but this is likely due to the relatively later advent of the Internet and e-mail.

The second wave of technological innovation came as computers displaced interviewers in favor of self-administered questionnaires (SAQ). Even in the context of face-to-face interviewing, respondents frequently interact directly with the computer when providing sensitive information. Considerable evidence indicates that this approach leads to more accurate reports of sensitive information.

Both research and practice have demonstrated that computerization provides a huge advantage across modes, enabling researchers to implement much more complex questionnaires while reducing respondent and interviewer burden. Automating the route that each respondent takes through the survey ensures that researchers are able to ask questions that are relevant to each respondent without the interviewer or respondent needing to figure out how to follow complex skip instructions (for questions that branch from core questions). Thus, computerization reduces the burden on both interviewers and respondents while providing relevant data for researchers.

Web surveys became popular as Internet coverage improved. Sometimes these surveys were adjuncts to traditional survey modes where, for example, a mail survey might invite the respondent to provide their answers online instead of on paper. Web-only survey data collection has a considerable history of controversy given its close association with non-probability sampling. Thus, none of the major NSF-funded surveys (and few government-sponsored surveys more generally) are web-only.

Data collection decisions often imply other survey design decisions. For example, a particular method of data collection is typically yoked with a specific sampling approach; thus, most CATI surveys are conducted with random-digit dial (RDD) samples. Further, with national face-to-face designs, cost constraints make clustered designs necessary; thus, face-to-face interviews are often done with area probability samples. So the mode of data collection and the sampling frame are typically bundled in a package. Thus, the mode decision implies bundles of survey features, and these influence the entire spectrum of survey error sources, as well as the survey's cost and timeliness:

Non-observation error:
- Coverage (since each mode is linked with a sampling frame and access)
- Nonresponse
- Random sampling error (clustering, stratification, sample size)
- Sampling bias (e.g., with non-probability web panels)

Observation errors:
- Random measurement error
- Response order effects (primacy/recency)
- Interviewer effects on respondent reports (none in mail, some in telephone, many in face-to-face)
- Social desirability bias (the tendency of respondents to provide inaccurate reports to make themselves appear in a more favorable light)

Given the central importance of the choice of survey mode, this choice may reflect a number of important features of the survey. The first and foremost of these features is cost; face-to-face interviewing is extremely expensive, but often organizations will sacrifice sample size for the

higher response rate and better data quality that are often associated with in-person surveys. Second, different sampling frames have their own particular issues. With face-to-face interviews, the most common approach is area sampling due to the need to cluster the sample geographically. With telephone, list-assisted frames of telephone numbers are most common, and with self-administered surveys, the frame determines whether the administration mode is mail or web, although the U. S. Postal Service's address list is coming to be used for both. Generally, web surveys are hampered by the lack of suitable sampling frames, which has prevented the bundling of mode and frame since no standard approach to sampling Internet users has yet emerged.

Coverage error is often the second most important consideration when selecting a mode of data collection. This error arises when not every unit in the population is represented on the sampling frame. If there are differences between the units who are on the frame and those who are omitted, then coverage error becomes a problem. For example, web surveys exclude those who do not have Internet access, and landline-only telephone surveys exclude those who only have cell phones. Considerable evidence indicates that coverage error is a significant concern for web surveys. This coverage error manifests itself in the "digital divide", the large number of substantial demographic and non-demographic differences between people with Internet access and those without.

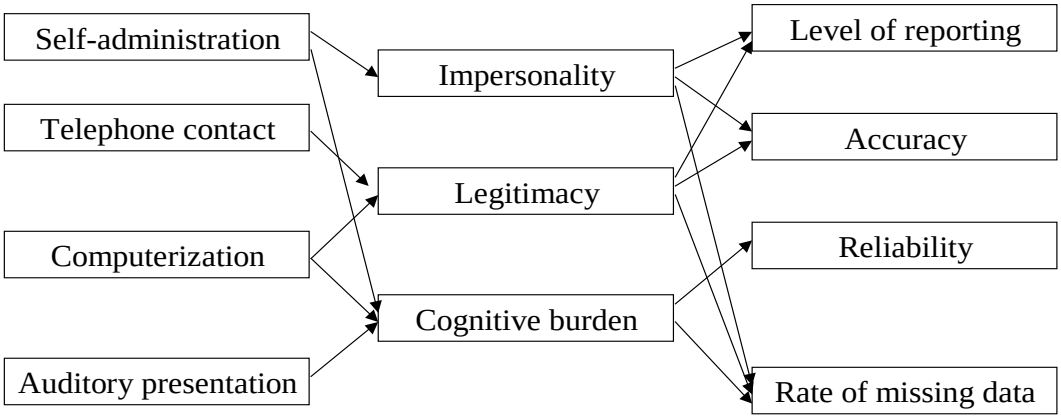
The key statistical consequences of non-observation error (encompassing sampling error, coverage error, and nonresponse) are inflated variance and bias. Unadjusted estimates, such as means or proportions, from non-probability samples are likely to be biased estimates. Similarly, estimates from any sample affected by nonresponse bias or coverage bias will, by definition, produce at least some biased estimates. The size and direction of the bias depend on two factors: one reflecting the proportion of the population with no chance of inclusion in the sample, and the second reflecting differences in the inclusion probabilities among different members of the sample who could in principle complete the survey.

Measurement error is also influenced by mode. Couper (Chapter 5, in Groves et al., 2005) has proposed five features that should be considered in understanding the impact of mode on measurement error:

1. The degree of interviewer involvement (e.g., mail and web feature low levels, CAPI high levels)
2. The degree of interaction with the respondent (e.g., eliciting opinions from a respondent vs. abstracting information from respondent records)
3. The degree of privacy (e.g., presence of interviewer, third parties, etc.)
4. Channels of communication (e.g., how questions are communicated to respondents and how they respond)
5. Technology use (paper vs. computer).

Models of this type can be thought of as proposing mechanisms through which differences in mode may result in differences in the data. Similarly, Tourangeau and Smith (1996) have developed a model for how the data collection mode can affect various features of the data; the version below was adapted by Tourangeau et al. (2000).

Model of the Impact of Data Collection Mode



Source: Tourangeau, Rips, and Rasinski (2000, Figure 10.1)

This model implicates three psychological variables that mediate the association between data collection mode and differences in the resulting data. The first is impersonality, which is how personal or impersonal the respondent perceives the survey interaction and context to be. The second is legitimacy, which is whether or not the survey and/or sponsor seem legitimate to the respondent. And the third variable is the cognitive burden of the survey response process for the respondent.

Within self-administered modes, there has been considerable discussion about potential mode differences and some research has been conducted examining these potential differences. In a meta-analysis of these studies, Tourangeau and Yan (2007) found that there is no significant effect of computerization on response. This is good news for researchers because it means that they can take advantage of the full range of self-administration options without great concern about differences in the resulting data.

Another important area of data collection mode research has been on the use of mixed-mode designs. These often represent best practices with regard to reducing costs and improving response rates. For example, surveys may begin with a less expensive mode of data collection (to reduce cost) and then switch to more expensive modes (to reduce non-response). The last few decennial censuses in the U.S. have followed this model, starting will mail and following up with mail nonrespondents using face-to-face data collection. There are a number of other ways that these mixed mode designs have been done. In some cases, one mode may be used for sampling and recruitment and another for data collection; for example, mail may be used to sample and recruit people to participate in a web survey. Other surveys have used one mode at the start of a panel survey and then changed modes in later waves; for example, the Panel Survey of Income Dynamics (PSID) used face-to-face interviewing for the initial wave and then switched to

telephone thereafter. Another design that has received considerable research attention uses different modes for different segments of the population, for example, web surveys for those with Internet access and mail for those without. Sometimes one mode will be used to follow-up for another mode, as with the decennial censuses. In this case, data collection often begins with a cheap mode, such as mail, to recruit willing respondents and then switches to a more expensive mode, such as telephone, to recruit reluctant respondents. Longitudinal surveys may reverse this process, starting with the expensive, high response rate mode first to maximize recruitment and then transitioning to a less expensive mode for later waves. Less common are approaches that implement different modes to collect different types of data from the same respondents.

One goal for some surveys using mixed-mode designs is maximizing comparability between modes; this implies that the same person should theoretically (or in practice) provide the same responses to a survey conducted by any mode. This has brought about a design approach known as unimode designs. The notion behind the unimode design is that mode effects should be minimized at all costs. When implementing a unimode design, there are a number of common considerations that arise that are not reflected in single-mode designs. For example, instead of optimizing the survey features for a single mode, questionnaires and procedures would need to be designed to ensure equivalence across modes. If mail is one of the modes used, then any computerized mode should attempt to mimic the mail design and flow as closely as possible. This means that researchers are unable to take advantage of many of the design features that are enabled by computerization, such as complex skip patterns and branching. Likewise, show-cards should not be used in a face-to-face mode if that mode will be paired with a telephone survey or some other method in which the respondent cannot be presented with the show card.

However, not all researchers agree that this is the best way to conceive of mode effects, particularly when maximizing comparability between modes is a secondary concern to minimizing total error. This alternative way of thinking conceptualizes mode effects as differential measurement error. One model for better understanding the differential measurement error framework, suggested by Tourangeau (Tourangeau et al., 2013), is:

$$wb_A + (1 - w)b_B$$

Where b is a measurement effect, A and B are different modes, and w is the average bias.

Using this model, it becomes clear that making the error in A match the error in B may not result in the lowest amount of total error. Instead, error in both modes should be minimized in order to minimize overall error. This is a different objective from the one adopted by the unimode approach; the goal is not necessarily maximizing comparability but minimizing overall error. So, under this approach, a researcher using telephone and face-to-face surveys would certainly want to use a show card in the face-to-face survey if it might reduce measurement error in that mode, even if there is no analogue to the show card in the telephone mode.

Best practices in data collection are reasonably well understood in the context of traditional survey methods. However, there may be opportunities or challenges that still need to be addressed. This is particularly true when implementing newer approaches to data collection such

as cell-phone surveys and surveys on mobile devices or when trying to balance comparability with minimum error in the mixed-mode context.

While differences between data collection modes have generated a lot of attention in research there is still more research that needs to be conducted to further develop our understanding of mode effects. Measuring mode effects is important but can be costly, meaning that it is not regularly done outside of the academic context. While some work has been done on developing models to parse non-observation errors from observation errors in mode effects studies, more research is warranted in this area. Future research should also take advantage of opportunities to compare with gold standards such as administrative records and make greater use of within-subject designs. Below is a list of proposed data collection mode research topics that the NSF should consider funding.

Areas for future research:

- Funding research aimed at developing methods and best practices for optimizing data collection across modes rather than mode comparison studies that are simply descriptive
- Mail/Address-Based Sampling vs. Telephone/Random Digit Dialing in the changing landscape of falling response rates for telephone surveys
- Minimizing measurement error vs. unimode designs for mixed mode studies
- Disentangling observation and non-observation differences in mode comparisons
- Reducing measurement error in self-administered surveys
- Identifying challenges and opportunities in new and changing modes such as cell phones, tablets, and other mobile devices

References:

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. New York: John Wiley & Sons. (Second Edition, 2009)

Tourangeau, R., Conrad, F.G., and Couper M.P. (2013). *The Science of Web Surveys*. New York, NY: Oxford University Press.

Tourangeau, R., Rips, L.J., and Rasinski, K.A. (2000). *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Tourangeau, R. and Smith, T.W. (1996). Asking Sensitive Questions: The Impact of Data Collection, Mode, Question Format, and Question Context. *Public Opinion Quarterly, 60*(2), 275-304.

Tourangeau, R. and Yan, T. (2007). Sensitive Questions in Surveys. *Psychological Bulletin, 133*, 859–883.

The Use and Effects of Incentives in Surveys – Eleanor Singer

Twenty years ago, there was a consensus that incentives should not be used for surveys that were less than an hour in length. There was also great debate about whether response rates to some of the large national household surveys were declining or not. Today there is no doubt that response rates are declining even for the best-managed, costliest, most important surveys, and incentives are used in most of them. In the ANES, PSID, and GSS, which are the largest surveys funded by NSF, the largest portion of nonresponse is attributable to refusals rather than non-contacts. Monetary incentives, especially pre-paid incentives, are capable of reducing nonresponse, primarily through reducing refusals. However, very little is known about the effects of incentives on nonresponse bias, which signals an important area of future research that the NSF should consider funding.

Survey research has expended considerable effort and research funds in examining the effects of incentives on a variety of outcomes. These outcomes include response rates in different types of surveys, sample composition, response quality, and response distributions. Much of this research has been conducted in the context of experiments attempting to improve response rates. The findings presented in this section may not apply uniformly to all surveys. The surveys that these findings are most applicable to are:

- Large, usually national surveys done for purposes related to social research
- Often longitudinal
- Typically funded by government statistical agencies or research organizations supported with government research grants
- Results are intended to be generalizable to a defined population

Market research, customer satisfaction surveys, polls with a field period of a week or less, and similar surveys are not included, and the findings presented below may not apply.

One of the most consistent findings of research on survey incentives has been that pre-paid incentives increase response rates; this has been demonstrated across survey modes. In the cross-sectional mail mode, a meta-analysis (Church, 1993) found that prepaid incentives yielded significantly higher response rates than promised or no incentives, monetary incentives yielded higher response rates than other gifts, and response rates increased with increasing amounts of money, though not necessarily linearly. Edwards and colleagues (2005) reported very similar results in a subsequent meta-analysis, and with very few exceptions, more recent experiments have yielded similar findings.

Two meta-analyses of experiments with interviewer-mediated surveys (Singer et al., 1999; Cantor et al., 2008) found that while the results were generally similar to those in mail surveys, the effects of the incentives were generally smaller. More specifically, Cantor and his colleagues present a number of findings regarding incentives in interviewer-mediated surveys:

- Prepayment of $1 to $5 increased response rates from 2-12 percentage points over no incentives
- Larger incentives led to higher response rates, but at a decreasing rate
- The effect of incentives has not declined over time, but baseline response rates have dropped substantially
- Prepaid incentives used during refusal conversion had about the same effect as those sent at initial contact, but at a lower cost
- Promised incentives of $5 and $25 did not increase response rates; larger incentives sometimes did
- More recent experiments involving interviewer-mediated surveys, including face-to-face surveys, have found similar patterns of results

Longitudinal surveys have typically made use of incentives as part of a larger motivational package designed to both recruit and retain respondents. As the cross-sectional context, incentives in longitudinal surveys increase response rates, usually by reducing refusals but occasionally by reducing non-contacts (McGrath, 2006). A considerable number of studies have indicated that an initial payment may continue to motivate respondent participation in subsequent waves, meaning that an up-front investment in incentives may have greater effect for longitudinal surveys than cross-sectional ones (Church, 1993; Singer and Kulka, 2002; McGrath, 2006; Creighton et al., 2007; Goldenberg et al., 2009). Other research has indicated that prepaid incentives in longitudinal surveys may increase response among those who have previously refused but not among those who have previously cooperated; this may indicate a "ceiling effect" (Edwards et al., 2005; Zagorsky and Rhoton, 2008). Further, a study by Jäckle and Lynn (2008) found that (1) incentives at multiple waves significantly reduced attrition in all waves, (2) the incentives did so proportionately among certain subgroups and so did not reduce attrition bias, (3) the effect of the incentive decreased across waves, and (4) incentives increased item nonresponse. The authors conclude there was a net gain in information.

Recently, some research has examined the effects of incentives on response quality. Typically, item nonresponse and the length of answers given to open-ended questions are used to measure response quality, but other measures to assess accuracy and reliability would be desirable. There are two alternative hypotheses about the effect of incentives on response quality. One posits that the respondent perspective is, "You paid me some money and I am going to do this survey, but I am not going to work very hard at it." The second hypothesis is that respondents feel they have an obligation to answer the survey and do their best to answer correctly. Most research has found little to no support for the notion that incentives influence response quality; only one study found that incentives increased item nonresponse across waves in a panel study but decreased unit nonresponse, resulting in a net gain of information (Jäckle and Lynn, 2007; McGrath, 2006). Cantor and his colleagues (2008) argue that the two hypotheses need to be tested, controlling for factors such as survey topic, survey size, type of incentive (e.g. prepaid, promised, refusal conversion), and whether studies are cross-sectional or longitudinal. For this, a much larger pool of studies would be required, and this is an area warranting future research.

1. Medway (2012) examined this question using a very large pool of measures of effort (e.g. item nonresponse, length of open-ended responses, straightlining, interview length, underreporting to filter questions, lack of attention to question wording, use of round

numbers, order effects, etc.) as well as the potential interaction of a number of demographic characteristics with receipt of an incentive. The findings of this study indicated that there were significant differences on only two effort indicators - reduced item nonresponse and time to complete; neither was significant once cognitive ability and conscientiousness were controlled. There were also no significant interaction effects between demographics and incentives on an index of satisficing. But because this study was implemented using a telephone survey, an important research question that remains is whether or not the same results would be found in a self-administered survey context. The study by Jäckle and Lynn (2008) found greater effects of incentives on unit and item nonresponse in mail than in phone administration of same survey, indicating that the potential interaction effect of incentives and mode of data collection on data quality needs further research. Additional aspects of quality, such as the potential effects of incentives on reliability and validity, also need study. Some further areas in need of research are discussed in the paragraphs that follow. Incentives have been shown to influence sample composition, meaning that the characteristics of people recruited are altered when incentives are used. Some of the characteristics of the sample that have demonstrated differences in response to incentives are: education, political party membership, socioeconomic status, and civic duty. However, the majority of studies reporting these findings have done so as *ex post facto* explanations. Specific attempts to use incentives to improve response rates among certain categories of respondents who may be less disposed to respond because of their lower interest in the survey topic have received only qualified support. Importantly, no studies have looked at the effect of incentives targeted to refusals. Theoretically, one would expect such targeted incentives to be more successful in changing the composition of the sample, thereby potentially reducing nonresponse bias, so this is an area ripe for future research.

2. Another aspect of incentives that has generated some controversy is that of differential incentives. Differential incentives refer primarily to refusal conversion payments, which are typically higher than prepaid incentives. Two arguments have been made in favor of differential incentives. First, they are more economical than prepaid incentives, and second, they are more effective in reducing bias. Historically, the primary argument against using differential incentives is that they are unfair. However, economists argue that differential payments are fair; those who refuse consider the survey more burdensome and therefore need/are entitled to bigger incentives. Respondents who are informed about differential incentives consider them unfair but say they would respond to a new survey by same organization even when told it engages in practice (Singer et al., 1999a). Experimental research indicates that these respondents do indeed respond to surveys purportedly by another organization a year later; there is no statistically significant difference in response due to receipt of an incentive or perception of fairness. The research on differential incentives has generated two recommendations for best practice. First, survey organizations should offer small, prepaid, incentives to all sample members; this will increase sample size and help satisfy the fairness criterion. Second, they should offer differential incentives to those who refuse (or a subsample) for bias-reduction reasons, but this practice should be accompanied by research to detect whether or not refusal conversion actually reduces bias.

3. To maximize the value and return from incentives, pretesting is extremely helpful. Different people may be motivated by different appeals; research is needed to find out which are most effective for a particular study. This is true at the individual-study level and in a more general sense across survey research. Researchers should also test the effectiveness of different combinations of appeals in introductory materials including, but not limited to, monetary incentives. For large and expensive surveys, a pretest that can yield quantitative estimates of likely response and the effectiveness of incentives, by important subgroups, may be warranted. Researchers should also take care to use pretesting to investigate respondents' and nonrespondents' perceptions of the costs and benefits of survey participation. The goal of such research is to develop empirically based efforts to improve the survey experience; incentives are a part of this equation, but the net benefits extend well beyond simply informing how to best spend incentive money.

4. More research is needed on how best to use incentives to bring about decreases in nonresponse bias for the most important dependent variables in a survey. Since all prior studies have used prepaid incentives, one recommendation is to focus research on targeted refusal conversion payments instead or in addition. Another recommendation for future research is to explore using address-based sampling rather than RDD to draw the initial sample for telephone surveys, sending letters with prepayment to a random subsample, and measuring nonresponse and nonresponse bias in both groups. A number of studies have shown that advance letters including incentives can substantially increase response in telephone surveys (letters without incentives do not appear to have such effects). However, the percentage of RDD sample members for whom addresses can be obtained is limited, and they tend to differ from those for whom addresses cannot be obtained. As a result, this tactic results in recruiting more respondents like those who would have been recruited even without the letters (Curtin et al., 2005), thus minimally affecting nonresponse bias.

5. More research is needed on how best to use incentives to bring about decreases in nonresponse bias for the most important dependent variables in a survey. Since all prior studies have used prepaid incentives, one recommendation is to focus research on targeted refusal conversion payments instead or in addition. Another recommendation for future research is to explore using address-based sampling rather than RDD to draw the initial sample for telephone surveys, sending letters with prepayment to a random subsample, and measuring nonresponse and nonresponse bias in both groups. A number of studies have shown that advance letters including incentives can substantially increase response in telephone surveys (letters without incentives do not appear to have such effects). However, the percentage of RDD sample members for whom addresses can be obtained is limited, and they tend to differ from those for whom addresses cannot be obtained. As a result, this tactic results in recruiting more respondents like those who would have been recruited even without the letters (Curtin et al., 2005), thus minimally affecting nonresponse bias.

6. Another important area for future research should measure long-term effects of incentives on public willingness to participate in research going forward by adding questions about expectations to a sample of existing cross-sectional surveys (e.g., GSS, Surveys of

Consumers). There is no evidence that the increasing use of incentives has had long-term effects on such willingness, but existing studies have looked at change over short intervals and with panel respondents, who may consider multiple waves as one survey.

7. Additional future research is also needed to examine changing interviewer expectations about the use of incentives and the effect of these on their response rates. It is plausible to assume that interviewers' expectations will change over the long run as a result of their experience with the increasing use of incentives. The decline in response rates over the past 15 years may in part reflect changing interviewer expectations and behavior, cultivated by reliance on monetary incentives. To shed light on whether and how motivations for survey participation are changing, it would be useful to sponsor systematic inquiry over time into reasons for responding and not responding, using experiments and open-ended questions. Do motives differ by age, gender, ethnicity, race, and income? Are altruistic motives declining?

8. There is no good evidence that monetary incentives reduce response rates, but there are indications that there may be ceiling effects (Singer et al., 1999; Zagorsky and Rhoton, 2008). Why should this be? Why are incentives not simply additive with other motives for responding?

9. Research is also needed to find out if incentives are coercive. Do they have undue influence on sample members' decisions about survey participation, in the sense of inducing them to undertake risks they would not otherwise take? Research so far suggests it does not, but experiments are needed that employ a wider range of incentives and greater variety of risks among differentially susceptible populations.

10. Finally, research is needed on the cost-effectiveness of incentives compared with other efforts to increase response rates and reduce nonresponse bias.

To summarize, here are some areas for future research:

- Using incentives to reduce nonresponse bias – are targeted refusal conversion payments effective?
- Using ABS instead of RDD to draw the initial sample for telephone surveys so that a prepaid or refusal incentive can be sent
- Understanding respondents' and nonrespondents' perceptions of the costs and benefits of participation and how incentives factor in
- Are there ceiling effects for incentive use in relation to response rates?
- Experimental pretests of introductory survey materials
- More experimental research on the potential of incentives to affect response quality
- Long-range effects of incentives on the public's willingness to participate in surveys
- Long-range effects of incentives on interviewers' expectations and their response rates
- Are incentives coercive?

References:

Cantor, D., O'Hare, B., and O'Connor, K. (2008). The Use of Monetary Incentives to Reduce Non-Response in Random Digit Dial Telephone Surveys. In: Lepkowski, J.M., Tucker, C., Brick, J.M., de Leeuw, E.D., Japec, L., Lavrakas, P.J., Link, M.W., and Sangster, R.L. (Eds.), *Advances in Telephone Survey Methodology*, (pp. 471-498). New York, NY: John Wiley & Sons, Inc.

Church, A.H. (1993). Estimating the Effect of Incentives on Mail Survey Response Rates: A Meta-Analysis. *Public Opinion Quarterly, 57*(1), 62-79.

Creighton, K.P., King, K.E., and Martin, E.A. (2007). The Use of Monetary Incentives in Census Bureau Longitudinal Surveys. *Survey Methodology Research Report Series no. 2007-2*. Washington, DC: U.S. Census Bureau.

Curtin, R, Stanley Presser, and Eleanor Singer. (2005). Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly 69*(1): 87–98.

Edwards, P, Cooper, R., Roberts, I., and Frost, C. (2005). Meta-Analysis of Randomised Trials of Monetary Incentives and Response to Mailed Questionnaires. *Journal of Epidemiology and Community Health 59*(11): 987–999.

Goldenberg, K.L., McGrath, D., and Tan, L. (2009). *The Effects of Incentives on the Consumer Expenditure Interview Survey*. Washington, DC: U.S. Bureau of Labor Statistics, internal report, 5985–5999.

Jäckle, A. and Lynn, P. (2008). Respondent Incentives in a Multi-Mode Panel Survey: Cumulative Effects on Nonresponse and Bias. *Survey Methodology, 34*, 105–117.

McGrath, D. (2006). An Incentives Experiment in the U.S. Consumer Expenditure Quarterly Survey. In: *Proceedings of the Survey Research Methods Section,* (pp. 3411–3418). Alexandria, VA: American Statistical Association.

Medway, R.L. (2012). Beyond Response Rates: The Effect of Prepaid Incentives on Measurement Error. Dissertation available at http://hdl.handle.net/1903/13646.

Singer, E., Groves, R., and Corning, A. (1999). Differential Incentives: Beliefs about Practices, Perceptions of Equity, and Effects on Survey Participation. *Public Opinion Quarterly, 63*(2), 251-260.

Singer, E. and Kulka, R.A. (2002). Paying Respondents for Survey Participation. In: Ploeg, M.V., Mott, R.A., and Citro, C.F. (Eds.), *Studies of Welfare Populations: Data Collections and Research Issues*, (pp. 105-128). Washington, DC: National Academies Press.

Zagorsky, J.L. and Rhoton, P. (2008). The Effects of Promised Monetary Incentives on Attrition in a Long-Term Panel Survey. *Public Opinion Quarterly, 72*, 502–13.

Building Household Rosters Sensibly – Kathleen T. Ashenfelter

Many probability sample surveys begin with a series of questions intended to elicit rosters of all household members at an address so that a random selection can be made among these members. Other surveys use roster techniques as the basis for determining which household members need to answer the survey questions and which do not qualify as respondents, while still others generate rosters in order to help respondents more accurately count the people living at the sample address. This paper generalizes across surveys with different goals but makes the case that an accurate roster is necessary for most survey goals.

Commonly-used rostering approaches, usually based on who "usually" lives or stays at a residence, are generally assumed to be effective for their purpose, but they have some problems. The rules used to determine who should be considered a household member and who should not are remarkably unstandardized across U.S Census Bureau surveys and most other large surveys that utilize these procedures. For example, the Census Bureau employs different rules in different surveys. Even within a single survey, different instructions for interviewers and respondents sometimes contradict one another. However, these inconsistencies have often been warranted due to the differing goals among surveys. In some cases, differences among rules for determining who should be counted as a resident in a household may be sensible depending on the purpose of or the constraints faced by certain surveys. Thus, household rostering is an arena in which new conceptual and operational research is warranted and could help survey researchers to optimize a procedure that is used across a large number of surveys and for the decennial Census.

While determining how many people are living at a sample address seems like it should be a fairly straightforward and simple task, there are a number of problems that survey designers must anticipate when formulating these rules. For example, some people may not live at a particular address all the time, such as retirees who spend summers and winters living in different states, or college students who may live at school for part or most of the year.

Most research on rostering has been based on the methodology used by four major surveys conducted by the U.S. Census Bureau: the Decennial Census, the American Community Survey (ACS), the Survey of Income and Program Participation (SIPP), and the Current Population Survey (CPS). Although these surveys are all presumably designed to ascertain equivalent lists of all people living at each selected address (with each person being assigned to only one address to prevent double-counting in principle or in practice), these four surveys employ different procedures and, therefore, seem likely to yield different results. More importantly, each individual survey occasionally offers instructions to respondents and interviewers that are logically inconsistent with one another, and the instructions sometimes entail the use of terms and concepts that are not sufficiently or clearly defined to allow respondents to easily and uniformly interpret the instructions and then comply with them whilst generating their responses. Research is needed that directly compares the outcomes from different approaches to generating a household roster so that we can assess whether or not different approaches aimed at achieving the same result are actually effective or whether key differences arise in the rosters drawn under these different methods.

One of the major issues that has been identified with current practice in household rostering is the presence of inconsistencies with respect to the date or dates used as temporal reference points for determining whom should be listed as a resident. For example, for the 2010 Decennial Census, respondents who were having trouble filling out the paper form had the option to call in and provide their responses to the Census over the telephone. Some instructions for interviewers tell them to count people who were residing in the household **on** April 1, 2010, such as: "*count the people living in this house, apartment or mobile home on April 1, 2010.*" But then in another portion of the same set of instructions, the language changes and becomes less specific, instructing the interviewer, for example to "*indicate foster children if they were staying at the address on or around April 1.*" In the latter example, the modification to include dates **around** the target date is a significant departure from the original instruction. The ACS contains similar inconsistencies in its set of instructions for generating the roster. In one section, the instruction to respondents says, *"Include everyone who is living or staying here for more than two months, include yourself if you are living here for more than two months, including anyone else staying here who does not have anywhere else to stay, even if they are here for less than two months."* Thus, in a single section, the instructions provide contradictory directions to the respondents for which people they should include on the roster.

Another common area of roster-related ambiguity involves how a survey should enumerate college students and military personnel. For example, the 2010 Decennial Census instructions indicated that: "*college students and armed forces personnel should be listed where they live and sleep most of the time*." But "most of the time" is never defined, leaving the respondent (and/or interviewer depending on the mode of administration) to arbitrarily determine who should be counted. Similarly, in the 2011 ACS, interviewers using the Computer Assisted Personal Interview (CAPI) mode were directed to read the following sequential list of instructions and questions to respondents:

- I am going to be asking some questions about everyone who is **living or staying** at this address. First let's create a list of the people, starting with you.
- Are any of these people away NOW for more than two months, like a college student or someone **living** in the military?
- Is there anyone else **staying** here even for a short time, such as a friend or relative?
- Do any of these people have some other place where they **usually stay**?

The instructions above give no temporal reference point or any definitions for what constitutes these rather vague concepts of time period. That is, there is no clearly defined set of time-based metrics that the interviewer or respondent can use to determine what a "short time" or "usually stay" means. These terms could mean very different things to many respondents to a particular survey, leading to differences in the final inclusion or exclusion of individuals in the resulting household roster.

Other problems exist beyond inconsistency issues. One example is that instructions to respondents about whom to count and whom to exclude are often vague. Additionally, some survey instructions are intentionally designed as a feature of the instrument that is only seen by interviewers and never shown to respondents during the survey interview. From a

methodological standpoint, this asymmetry in availability of rostering information could impact data quality. From a human factors and usability standpoint, the additional context found in these interviewer instructions could be extremely helpful for respondents while they are answering the roster questions. Another common issue is that household rostering procedures are often unnecessarily complicated and include complicated branching patterns, which increases the opportunity for mistakes.

Roster complexity is an important, although often overlooked, contributor to the relative ease of use of a survey instrument, and is a concept that warrants in-depth research. American households and living situations can be very complex. Rostering rules typically attempt to account for this complexity by providing instructions to interviewers and respondents for how to accurately determine whom to actually count as a member of the household. There are many living situations that increase the difficulty of building household rosters accurately according to the given set of residence rules, including the following common issues, which do not reflect a complete set of the diverse circumstances represented across American households:

Complex Households

- Large households, which may or may not include members of extended families;
- Tenuous attachment (Tourangeau, 1993);
- Roommates;
- Roomers and boarders (Hainer et al., 1988; McKay, 1992);
- Students who attend boarding school;
- College students;
- Commuters who may or may not have a second residence to be closer to their place of work Monday-Friday;
- Babies, whom certain respondents tend to exclude from household rosters;
- Children in a shared custody arrangement in which the children are not at the sample residence every day but might be considered as usually living or staying there by one or both parents;
- People temporarily living away from the sample address for a variety of reasons, either in their own second residence or a residence not owned by them;
- Concealed household members due to respondents' fear of losing their welfare benefits if the government discovers that the additional person or people usually live or stay at the sample address. Respondents who are immigrants, especially those containing household members who are illegally residing in the United States may also fear deportation, arrest, or other serious consequences that they believe are associated with becoming linked to an identifiable address or residence (Hainer et al., 1988; McKay, 1992);
- Homelessness, either temporary or permanent; and
- New categories, such as "couch surfers" who find residences where they can sleep on a couch, usually for a short time period, and who usually go online and use the Internet (e.g, Web sites such as Craigslist.com), to locate amenable couch-owning residences.

One common approach to addressing these challenges to accurate household rostering has been to use equally complex systems of rules, the goal of which is to determine who should count as a member of the household. However, a major drawback to this approach, especially for researchers hoping to compare data between surveys, is that these rules are not standardized in

terms of content or structure. The same lack of consistency can be found across surveys if one examines the definitions provided for important terms and concepts contained within the survey questions. For example, the concept of "usual residence" is ubiquitous in rostering questions and can seem like a relatively simple concept upon initial consideration. However, consider the wide variety of methods that are employed in the process of determining whether an address is the usual residence for a generic individual named Joe:

- Does Joe contribute money for rent, food, bills, or anything else?
- How frequently does Joe sleep here?
- Does Joe use this address to receive mail or phone messages?
- Does Joe usually eat here?
- Do you consider Joe a member of the household?
- Does Joe have another place/places where he usually stays?
- Does Joe perform chores like cleaning?
- Does Joe have a say in household rules?
- Does Joe usually live or stay here?

Compared to the large number of different ways that someone can be considered a member of the household, there is a proportionally small body of research that has examined whether complex living situations have a significant effect on response tendencies and on overall data quality. Although it is possible that incorporating complex rostering rules into the design of a survey is one solution to the challenges presented by complex households, there simply has not been enough research conducted in order to draw this conclusion. Many programs of extensive empirical research are sorely needed in order to inform survey designers' and researchers' approach to conducting household rostering.

Additional rostering topics that similarly want for further research include a line of experiments aimed at determining best practices for question branching and for identifying ways to reduce the cognitive and time-related burden, for both respondents and interviewers, associated with conducting or responding to interview questions that ask respondents to apply residence rules to generate some form of a roster. Additionally, more research on self-response rostering is also needed so that researchers may gain a clearer understanding about the impact that a rostering approach may have on the survey's data quality (instead of simply making assumptions about what the impact might be). Further, the opportunity to utilize a convergence of scientific evidence on which to base decisions about rostering approaches is absent from the corpus of survey methodology research. Specifically, much of the data that we do have about collecting roster-related data from hard-to-reach cases of highly complex living situations, and populations that were at high risk for Census under coverage, comes from a single survey, the Living Situation Survey, which was last conducted in 1993. Revisiting this targeted approach to understanding living situations in the present time period could provide a great deal of direction for designers of large surveys when they are determining how their surveys should approach household rostering.

Here is a brief summary of concepts that past work or current needs have indicated are main areas for future research:

- Identifying the effects of using different rules in different surveys
    - Even within the same survey, different instructions to interviewers and respondents sometimes contradict one another. What are the effects of this inconsistency?
    - Do different approaches yield different results in terms of accuracy? Response rate? Data quality? What specific impact does rostering approach have:
        - On response rates?
        - On data quality?
        - On interviewer and respondent satisfaction with the interview interaction itself?
        - On the respondents' opinion of the agency sponsoring the survey? Has it improved, declined, or stayed the same based on their experience with the interview? Does the direction or magnitude of the change seem to be related to the mode in which the survey was conducted?
- Finding whether the ways that surveys operationalize usual residence are consistent with respondent notions of usual residence
- Understanding which Basic Residence Rules make the most sense and are easiest to understand for respondents
- Determining the optimal wording for the rostering questions so that they are easy for the interviewer to consistently read, pronounce, and annunciate
- The literature lacks a set of strong conclusions based on large-scale, empirical investigations of *de jure* (e.g., a rule-based method for rostering meant to count a person where his or her legal residence is, regardless of where they happened to be staying when the survey was conducted) vs. *de facto* (e.g., location-based method based on where the person was at the time of the survey) types of residence rules
- Objectively assess the impact of newly-introduced technological tools and methods of analysis available online and on mobile devices for rostering purposes (overlays, pop-ups, etc.)

Ultimately, there are many avenues of research that have not been conducted to date related to the concept of household rostering. We need more information about what the ideal approach to the design, application, and presentation of Residence Rules might be and whether the same set of rules that makes sense to people also helps generate a more accurate head count.

References and further reading:

Cork, D.L. and Voss, P.R. (Eds.) (2006). *Once, Only Once, and in the Right Place: Residence Rule in the Decennial Census*. Washington, DC: National Academies Press.

Fein, D.J. and West, K.K. (1988). Towards a Theory of Coverage Error: An Exploratory Assessment of Data from the 1986 Los Angeles Test Census. *Proceedings of the Fourth Annual Research Conference*, (pp. 540-562). Washington, DC: U.S. Census Bureau.

Gerber, E.R. (1990). Calculating Residence: A Cognitive Approach to Household Membership Judgments among Low Income Blacks. Unpublished report submitted to the Bureau of the Census, Washington, DC.

Hainer, P., Hines, C., Martin, E. A., and Shapiro, G. (1988). Research on Improving Coverage in Household Surveys. *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, (pp. 513-539). Washington, DC: U.S. Census Bureau.

Martin, E.A. (1999). Who Knows Who Lives Here? Within-Household Disagreements as a Source of Survey Coverage Error. *Public Opinion Quarterly*, *63*(2), 220–236.

McKay, R.B. (1992). Cultural Factors Affecting Within Household Coverage and Proxy Reporting in Hispanic (Salvadoran) Households. A Pilot Study. Paper presented at the 1992 meetings of the American Statistical Association, Boston, MA.

Schwede, L. (1993). Household Composition and Census Coverage Improvement. Paper presented at the American Anthropological Association Annual Meetings, November 1993.

Schwede, L. and Ellis, Y (1994). Exploring Associations between Subjective and Objective Assessments of Household Membership. In: *Proceedings of the American Statistical Association Section on Survey Research Methods*, (pp. 325-330). Alexandria, VA: American Statistical Association.

Schwede, L., Blumberg, R. L., & Chan, A. L. (Eds.) (2005). *Complex Ethnic Households in America*. Lanham, MD: Rowman & Littlefield Publishers, Inc.

Tourangeau, R. (1993). Final Report: SIPP Roster Questions. Report Prepared Under Contract (No. 50-YABC-2-66023) with the Census Bureau by National Opinion Research Center. Washington, DC: Bureau of the Census.

Proxy Reporting – Curtiss Cobb

Not all answers to survey questions are provided by the sampled respondent; often the person selected to be the respondent in a survey is unavailable when the interviewer is at the home or on the phone. In these cases, some surveys will allow another person to respond on behalf of the target; this person is a proxy for the target. This proxy may be another member of the household such as spouse or child, or a friend or co-worker. Proxy reports, then, are the answers to survey questions about the respondent that are provided by someone other than the target respondent.

There are a number of benefits to using proxy reports. For example, they tend to allow faster data collection and are less expensive because fewer interviewer visits to the sampled household are required before getting a completed interview. Proxy reporting is also associated with increased cooperation rates. Many surveys collect proxy reports for things like political participation, immigration status, social stratification, employment status or changes, and health and illness. Thousands of research articles in the social sciences have been written based on data that include proxy reports, often without the express knowledge or acknowledgement of the authors. Important surveys that collect data from proxies include the Census, the Current Population Survey (CPS), the General Social Survey (GSS), and the National Longitudinal Study of Adolescent to Adult Health (Add Health).

In the CPS, the use of proxy reports has been estimated to save up to 17 percent on survey costs versus interviewing all household members separately. This savings does come at a cost in terms of data quality though: laboratory tests indicate that agreement between target reports and proxy reports may be as low as 67 percent for some types of questions, such as the number of hours the respondent worked in the previous week. Similar results have been found for items on voting turnout and other important variables. However, some other, more general, items have been demonstrated to have upwards of 92 percent agreement between proxies and targets in laboratory studies. Thus, future research may need to focus on the types of questions that are accurately provided by proxies and the types of questions that are not.

There are a number of potential risks associated with using proxy reports. They may be less accurate than target responses, which is a different type of error from the level of agreement that was mentioned above as both the target and the proxy could be wrong about some objective phenomenon. Proxy motivation may be different from target motivation, leading to differences in effort and data quality. There may also be fundamental differences based on the perspective of the response; the proxy is responding from the perception of another, whereas the target would be providing a self-perception. However, it is not clear that these concerns are uniform across question types and contexts, and in some cases proxy reports may be more accurate than target reports due to better perception of the object of the question or lower social desirability effects for sensitive questions. Again, more research is needed to inform best practices for types of questions that are best suited for proxy reports and those that are not.

A recent meta-analysis of 93 studies attempting to evaluate proxy reporting found that most of these studies lacked the basic features required to make accurate assessments of proxy reports (Cobb and Krosnick, 2009). Below is a list of the necessary features of a proxy study with the number of the 93 that included that feature in parentheses:

- Both Targets and Proxies should be interviewed (17)
- Targets and Proxies should constitute representative samples of the same population (25)
- Questions asked of Targets and Proxies should be identical (21)
- Independent, external measure of the attribute being assessed should be used to measure accuracy (76)

With these features, there are two research strategies that could be employed. In the first, sampled individuals are randomly assigned to provide answers for themselves or a proxy. Then the results are aggregated and compared. Because of random assignment, the results should be very similar, and these sets of results can then be compared against the external benchmark validation values. The second approach is to obtain measurements from matched targets and proxies. This approach allows researchers to evaluate the role of nonresponse and assess the association between reporting errors made by targets and reporting errors made by proxies.

In the meta-analysis of 93 studies, Cobb and Krosnick found that only six had the necessary features to evaluate proxy accuracy. It is important to note that accuracy is defined as a measure of validity and is different from level of agreement between the target and proxy. This finding indicates that much of the existing literature is inadequate for its intended purpose, and considerably more research is warranted despite the relatively large number of proxy studies that

already exist. Of the six studies actually suited to assessing proxy accuracy, there was substantial variation between studies in the accuracy of self-reports and proxy reports, but the two types of reports were equally accurate, meaning that self-reports and proxy reports tended to be accurate or inaccurate together, not divergently. This is a promising finding for the validity of using proxy reports because it appears that they can be relied upon. These findings need further research, particularly since five of the six studies are over 40 years old.

Despite not being adequately designed to assess proxy report accuracy, many of the other studies in the meta-analysis were still informative in other ways. A number of the relevant and important findings are summarized below:

- Cognitive studies reveal that memories about others are less elaborate, less experientially based, and less concerned with self-presentation
- Proxies anchor answers based on their own behaviors and attitudes
- Proxies estimate more versus recall
- Time together increases agreement and the similarity of cognitive strategies used to arrive at responses
- Proxies are more likely to under-report behavior and events, except those that involve care-taking activities
- Knowledge of and exposure to the question topic increases agreement
- Stable traits and characteristics lead to more target/proxy agreement than changing activities
- Observable information is easier for proxies to report on than unobservable information

While these findings suggest a number of best practices for using proxy reports in surveys, there are still a number of important areas for future research. First, there is a need for more correctly designed studies on the accuracy of proxy reports across a variety of topical domains. Second, more research is needed on identifying question designs that increase proxy accuracy. This involves identifying appropriate reference periods, question formats, and features that may increase proxy motivation. Third, future research needs to explore how the characteristics of proxies impact the accuracy of their reports; for example, are household members more or less accurate than non-household members as proxies? Lastly, researchers need to identify optimal strategies for implementing best practices when designing new questions that may be answered by either targets or proxies.

Areas for future research:

- Investigating the implications of increasing levels of proxy reporting
- Identifying questions that are and are not appropriate for proxy reports to be collected on
    - How accurate are proxy reports on different types of questions?
    - What causes proxy inaccuracy?
    - Under what conditions are inaccuracies most likely to occur?
- Understanding how questions can be optimally designed to maximize accuracy for both targets and proxies
- Investigating the impacts of question design features on proxy report accuracy
- Understanding the impact proxy characteristics have on the accuracy of their reports

Reference:

Cobb, C. and Krosnick, J.A. (2009). Experimental Test of the Accuracy of Proxy Reports Compared to Target Report with Third-Party Validity. Presented at the American Association for Public Opinion Research Annual Meeting, Hollywood, Florida.

Improving Question Design to Maximize Reliability and Validity – Jon A. Krosnick

Few topics in survey methodology have received more attention than questionnaire design, and there is a large and growing body of research findings that inform best practices. But for every questionnaire design topic for which there are accepted best practices, several topics generate a large amount of controversy, and these controversies need to be addressed by future research.

When evaluating questions and seeking to identify the best ways to ask questions, researchers should be mindful of three primary goals. The first is to minimize administration difficulty. That is, employ questions that can be asked and answered quickly and easily by respondents. Second, survey designers would like respondents to make as few completion errors possible. So if a respondent is asked to choose a point on a rating scale on a paper questionnaire, survey designers don't want them circling two or three points, saying, "I am somewhere in this range, but I don't know where." Lastly, all other things equal, researchers would like respondents to enjoy answering the question and not be very frustrated by it. But, all else is not equal. At the end of the day, researchers should be willing to use longer questionnaires and to accept some respondent frustration if that's what it takes to maximize the reliability and validity of the measurements. Fortunately, the literature suggests that what goes quickly and easily for respondents usually produces the most accurate data.

To understand how to optimize the design of a question, it is useful to understand the cognitive steps involved in question answering. In general, the process is thought to involve four steps:

1. Understand the intent of the question – that is, what information the researcher seeks, rather than literally what the words mean.
2. Search memory for relevant information
3. Integrate the relevant information into a summary judgment
4. Translate the judgment into the required format for expression

Engaging in each of these steps effortfully is known as "optimizing". Unfortunately, a growing body of research suggests that people sometimes do not perform all four steps before providing their response. Instead, people sometimes satisfice, employing cognitive shortcuts. This can occur in two ways. One is to superficially engage the two middle stages (searching and integrating) rather than doing so effortfully. This is what researchers call 'weak satisficing'. Alternatively, if the respondent has entirely given up on providing good responses, he or she may skip the middle two steps entirely and simply understand the question and then offer a response. This is called 'strong satisficing'. In this case, respondents look to the question and situation for cues pointing to apparently plausible answers that would be easy to justify without thinking.

A number of satisficing strategies may be employed by respondents, including:

- Selecting the first reasonable response
- Agreeing with assertions
- Non-differentiation in ratings
- Saying "don't know"
- Mental coin-flipping

Three primary causes of satisficing have been implicated by the existing research: 1) respondent ability, 2) respondent motivation, and 3) task difficulty. Existing research indicates that to the extent that researchers can enhance motivation and simplify the cognitive tasks required, respondents will be less likely to satisfice.

Another important perspective on questionnaire design considers the conversational norms and conventions that normally govern conversation. Survey interviews often conform to these norms and conventions but sometimes violate them. Questionnaires are essentially scripts for conversations between researchers and respondents, and respondents reasonably expect the same rules to govern these conversations as govern all conversations. But in fact, survey questionnaires often violate these rules, and as a result, respondents can be misled or confused. For example, imagine a conversation between Person A and Person B. Person A: "How are you today?" Person B: "Good". Person A: "And how's your day going?" This would be a violation of conversational norms, to ask the same question twice. And yet surveys often ask multiple questions to measure the same construct. Grice (1975) described a series of conversational norms, and surveys risk respondent frustration and confusion when such norms are violated.

Researchers often face a choice between measuring a construct with an open-ended question and measuring it with a closed-ended question. Open-ended questions have a number of distinguishing features:

- The question stems are asked identically of all respondents
- No response alternatives are suggested
- Verbatim transcription is required
- Interviewers may probe the respondent to say more about a topic
- Interviewers must be cognitively skilled
- Interviewers must be trained extensively
- Data must be coded into categories
- Analysis may be expensive and time-consuming
- Answers may be provided freely and without bias that could be caused by offering answer choices

Closed-ended questions have a number of distinguishing features:

- Both question stems and answer choices are standardized across respondents
- Respondents code their own answers into categories
- Interviewer training is simple

- Administration is fast and cheap
- Data are easy to analyze
- Results may be more objective: No bias from the questioner (in probing) or the coder

Decades of research comparing these question forms for measuring categorical variables with unlimited universes and numeric variables indicate that open questions yield more reliable and valid data than do closed questions.

One concern about open questions is that responses might be particularly susceptible to distortion due to salience. For example, if a survey asks, "What is the most important problem facing the country?", and the respondent happened to have seen a news story about crime on television the previous night, that may enhance the likelihood that the respondent would retrieve crime as a potential answer. In contrast, if asked a closed question offering a list of response options, that salience effect may be minimized. However, empirical evidence disconfirms this claim. In fact, salience appears to affect open and closed questions equally.

A third potential problem with open-ended questions if frame of reference effects. For example, if asked, "What is the most important problem facing the country?", the respondent must understand what counts as an acceptable answer. This can be ambiguous with open-ended questions, whereas the answer choices offered by a closed-ended question define the frame of reference of acceptable answers for respondents. Some studies have documented instances in which open-ended questions were problematically ambiguous in this regard, so wording should be designed to define the frame of reference as well as possible.

In contrast, studies have provided validation of a series of limitations of closed-ended questions. One concern is the notion that people without real opinions might be induced to express so-called "non-attitudes", answers that do not reflect real judgments in a person's mind. Studies seeking opinions about fictitious issues, for example, have documented such expressions. Second, if a question seeks a numeric answer and offers ranges (e.g., 0 to 2 hours, 3 to 6 hours, etc.), the ranges in the middle of the array are what respondents assume are most common and therefore attract respondents. Lastly, a closed-ended question offering an "other – please specify" option does not successfully elicit responses outside of the explicitly offered set – respondents feel compelled to choose among those options and rarely use the "other" category. When survey designers can't be sure of the universe of possible answers to a categorical question, an open-ended question seems to be the preferable route to go.

Rating scales are quite effective for placing respondents on continua, and the literature offers some guidelines about how to design them. A variety of principles can guide the decision about how many points to put on a scale. In order to learn as much as possible about respondents and to allow respondents to map their attribute onto the dimension most veridically, offering more points seems preferable. However, if too many points are offered, respondents might be uncertain about the precise meanings of the points, which would compromise reliability and validity.

Empirical evidence suggests that reliability and validity can be maximized by offering seven points on bipolar rating scales (with a zero point in the middle) and five points on a unipolar

rating scale (where the zero point is at the end). Branching bipolar dimensions improves data quality: first asking the respondent which side of the issue he or she is on (should military spending be increased, decreased, or kept the same?) and then measuring extremity (should it be increased a great deal, a moderate amount, or a little?).

When choosing labels for scale points, a series of goals are worth pursuing. First, respondents should find it easy to interpret the meanings of all the scale points. And respondents should conclude that the meanings of the scale points are clear. Third, all respondents should interpret the meanings of the scale points identically. Fourth, the scale point labels should differentiate respondents as much and as validly as possible. And lastly, the resulting scale should include points that correspond to all points on the underlying continuum.

Empirical research has shown that reliability and validity and respondent satisfaction are maximized by labeling all rating scale points with words, not numbers, and choosing words and phrases that have meanings that people perceive to be equally spaced across the continuum, with the end points phrased as extremely as possible. Labels should be brief phrases rather than sentences or paragraphs.

When selecting words to use in question stems or answer choices, textbooks have offered conventional wisdom without explicit empirical support for the assertions, such as:

- Choose simple, direct, comprehensible words
- Avoid jargon
- Be specific
- Avoid ambiguous words
- Avoid double-barreled questions
- Avoid negations
- Avoid leading questions
- Include filter questions rather than asking questions that do not apply to a respondent
- Be sure questions read smoothly aloud
- Avoid emotionally-charged words
- Avoid prestige names
- Allow for all possible responses

Empirical evidence supports some of these principles, and others remain to be tested adequately. Most importantly, the literature suggests that researchers will be advantaged in their word selection processes by software that draws on databases regarding the numbers of definitions of words, the familiarity of words, the complexity of words, status as homonyms and heteronyms, and other linguistic features to facilitate optimal selection.

These are just some of the areas in which the literature provides guidance regarding question design. Huge bodies of research suggest that agree/disagree, true/false, and implicit or explicit yes/no questions should never be asked (to avoid acquiescence bias), that non-opinion or unsure options should never be offered (to avoid failing to measure real opinions), and that questions asking respondents to recall the opinions they held at prior times or to explain the reasons for

their thoughts and actions should be avoided (because the answers people give usually lack validity).

Much more research is needed in this area to help guide question designers when making many other decisions on which empirical evidence does not yet exist. The evidence that does exist is summarized in some available publications (e.g., Krosnick and Presser 2010).

References:

Grice, H.P. (1975). Logic and Conversation. In: Davidson, D. and Harman, G. (Eds.), *The Logic of Grammar*, (pp. 64-75). Encino, CA: Dickenson.

Krosnick, J.A. and Presser, S. (2010). Question and questionnaire design. In: Wright, J.D. & Marsden, P.V. (Eds.), *Handbook of Survey Research*, (pp. 263-314). Bingley, UK: Emerald Group Publishing.

Perception of Visual Displays and Survey Navigation – Stephen Kosslyn

Survey research has always sought to understand the effects of questionnaire design. Existing research has examined a range of topics from particular design elements such as scale points, different types of response options (grids, drop-downs, thermometers, etc.), to the layout of questionnaires and inclusion of images on web surveys. This research has been and will continue to be extremely valuable for survey research. However, very little of this work has integrated these questions of survey design with the research from cognitive psychology on visual perception; this suggests an opportunity for future research that will integrate these fields and further the understanding of how to best design survey instruments.

The field of cognitive psychology has a very well developed understanding of visual perception and processing. This extends from the functioning of single neurons to the ways that different parts of the brain interact to incorporate and synthesize visual information and make it meaningful and also to match visual inputs with information stored in long-term memory to form associations. The canonical model can be simplified into three basic processes: 1) encoding, 2) working memory, and 3) long-term memory.

If encoding doesn't occur, then the information may as well not exist. There are three crucial potential bottlenecks: 1) detection (for example, if there isn't enough contrast between the edges it is impossible to see the edge), 2) organization, and 3) attention, meaning that most of what gets encoded and makes it to long-term memory are the bits of information that we pay attention to.

Working memory is important, because in order to encode with a high resolution we need to fixate on a stimulus in order to integrate information over time. Essentially what this means is that when we look at a stimulus, our eyes are moving imperceptibly several times per second, building a composite image of the stimulus in our working memory. There are two potential bottlenecks in this process. The first is limited capacity, meaning that there is only so much information that we can handle. The second is the need for visual change, which means that in order to register that information is being conveyed there has to be some visual change.

In the last step in the process, long-term memory must be accessed. In order to understand an information input, one must access information previously stored in long-term memory. There are three important potential bottlenecks at this stage. First, the incoming information must be clear so that connections can be made in long-term memory or it will be impossible for the people perceiving the message to make the correct connections in their long-term memory. Second, the appropriate knowledge must be accessed, meaning that the information being presented must be relevant to what the person already knows or the new information will be incomprehensible to them. Third, the surface information has to be compatible with the meaning; that is, the interpretation of the new information has to be compatible with the actual meaning of the information or there will be confusion.

From a survey design perspective, all of this implies that researchers should start with the information that needs to be conveyed and then work backward through the stages of the model. This means starting with the long-term memory and the connections that will be made in response to the new information input, then considering the working memory processes that will make sense of the new information, and lastly considering the stimulus that will be encoded.

Three psychological principles or cognitive communication rules are helpful to consider when conveying information and attempting to get people - respondents in the survey context - to access the correct information stored in long-term memory when they are presented with information, including requests for information, in the survey. The first rule is what Kosslyn (2007) calls the "Goldilocks Rule," which states that the detail of information being provided to respondents should be just right. Neither too much nor too little information should be given for the situation; respondents should get just enough information to make the correct connections between the survey question being asked and the existing information in long-term memory to report their response. This means that researchers need to know exactly what message they are trying to convey to respondents and what connections will be made in long-term memory when the response process is engaged so that the stimuli or survey questions in this context can be targeted toward exactly the right parts of long-term memory for respondents. Any unnecessary information provided may only complicate the process for respondents and reduce the chances of the correct information being accessed in long-term memory.

The second rule is what Kosslyn has termed the "Pied Piper Rule," which says that in order to make connections with a respondent's previous knowledge, researchers should lead them by using familiar and appealing information. This implies that researchers should have an intimate knowledge of who the respondents to the survey will be and adjust the background information, concepts, and language of the survey to match what they already know. From high-level concepts down to simple use of abbreviations, notation, or jargon, if survey respondents are able to connect with the survey and become interested, then the cognitive processes engaged during the response are more likely to be fast and accurate.

Kosslyn's third rule is what he called the "Judging a book by its cover rule," which calls for making the form of the message to respondents fit the meaning that they will infer and making this as simple as possible. For example, many people have experienced the Stroop Effect, where they are asked to ignore the words and state the color of the text for the following words: Blue,

Green, Yellow. This is a classic example of words and meaning coming into conflict, and it is something that survey researchers need to be careful to avoid. One promising area for future research that these three rules suggest is identifying optimal ways to use graphics and words to convey information on survey questionnaires. This is a powerful mode of communication because it allows researchers to enter information into memory using two forms of encoding, visual and verbal. The visual connotations may enable respondents to form connections more efficiently and effectively, which could improve responses. But this is an area of survey research that has received very little attention and thus needs more research.

For working memory, there are two important principles that can guide survey researchers as they move respondents from having encoded information, such as a request to answer a question, to processing it in working memory. The first principle is defined by what Kosslyn calls "The Rule of Four," which suggests that no more than four units or elements ever be displayed to respondents at one time. By simplifying the pages with which respondents interact, researchers can improve the chances that respondents are going to understand what they're seeing quickly and easily. By limiting the visual or verbal elements to four units, respondents will be much more able to process all of the information in working memory without exhausting its very limited capacity. Fortunately, each of the four units can also contain four units and this can trickle down several levels, but the key is that at no level should respondents be asked to process more than four units in their working memory. This implies that surveys should avoid using long lists and instead find ways to organize the elements of lists into clusters of no more than four elements that are similar and then drill down using sequential organization of these clusters.

The second principle is what Kosslyn calls the "Viva la Difference rule." It suggests that all new information should be indicated by change, such as changes in color or size of the stimuli. Similarly, this suggests that information that is not new should not change in how it is displayed. For example, elements of a survey that are being introduced to the respondent for the first time should be presented in a way that is different from everything preceding it so that the change can alert respondents to the new information.

Continuing to work backward through the visual perception process, we finally arrive at encoding. For encoding processes, there are three psychological principles that can guide survey researchers, the first of which Kosslyn terms the "Mr. Magoo rule." This rule suggests that information must be easily distinguished and recognized; at a very basic level this means that patterns need to be discriminable, and this in the survey context this implicates text patterns and response scales. For example, survey designers should avoid using all upper case or italic typefaces because they are less discriminable and thus harder for respondents to encode. Similarly, words should not be underlined for emphasis because it cuts off descending letters like 'g' or 'p', which is another important cue that is used in reading. When emphasis does need to be made, it should be sparing, and bold typefaces and colors should be used instead. Emphasis should be sparing because it is important to maintain the contrast between more and less important information. If all information being presented is equally important, then no emphasis should be used. Other important design features for researchers to consider include avoiding busy backgrounds and ensuring sufficient contrast in any visual display so that information does not end up being camouflaged to respondents.

The second rule of encoding is what Kosslyn calls the "birds-of-a-feather rule," and this rule stipulates that perceptual grouping laws should be used to organize visual features to ease the encoding process for respondents. The two important perceptual grouping laws that researchers should be aware of are 1) similarity and 2) proximity. Respondents will assume that things that are similar should be grouped together, and the same is true for things that are close together. Researchers need to be aware of this and take it into consideration when laying out the visual elements of a survey questionnaire.

The third rule is the "Rudolph the red-nosed reindeer rule," which is based on the principle that people notice what's different, and hence more important elements should be larger, brighter, moving, or more distinctively colored than less important elements. People can't help but pay attention to where change is because the human brain is a very good difference detector. Attention is grabbed by changes in size, color, and location, which is the biggest change. So survey designers should be sure to use differences and changes judiciously so that they draw attention to only the most important features of the questionnaire.

Some things that survey designers should be careful to avoid from a visual perception and processing perspective are meaningless animations (in web surveys) and any movement of questionnaire elements that are meaningless, because they will automatically attract attention and distract respondents from their primary task.

Areas for future research:

- Understanding how principles of visual perception influence respondents and data quality
- Understanding the data quality implications of taking advantage of display options available on web surveys to manipulate respondent visual processing
- Using visual perception rules to improve CATI interfaces for interviewers
- Determining best practices for survey design from a visual perception perspective

Reference:

Kosslyn, S. M. (2007). *Clear and to the Point: 8 Psychological Principles for Compelling PowerPoint Presentations*. New York, NY: Oxford University Press.

Cognitive Evaluation of Survey Instruments: State of the Science and Future Directions – Gordon Willis

Survey pretesting and cognitive interviewing in particular are important features of survey development, and contributions from psychology have made cognitive interviewing one of the most powerful and widely used pretesting tools in survey research. However, widely accepted best practices for cognitive interviewing have been elusive, and much of what is done in practice tends to be idiosyncratic. This provides an extremely important area for future research.

In the context of total survey error, response error as a form of measurement error is a type that researchers can control through either question or questionnaire design. The notion is that

because this is a serious, yet controllable type of error, it is worthy of attention and continued research, due to the fact that small changes in question wording and questionnaire design and format can make a substantial difference in the answers that respondents provide to questions. For example, simply asking respondents how much time they spend on a common daily activity often results in over-reports of that activity when compared to a questionnaire that first asks if the respondents engage at all in the activity and then following up to request the amount of time only respondents who report the behavior.

Cognitive testing is an applied approach to identifying problems, like those described above, in survey questionnaires, questions, and related materials, with the goal of reducing the associated response errors. Typically, a preliminary version of the questionnaire is developed, members of the targeted population are recruited and paid for their time, and then one-on-one interviews are conducted, usually in a face-to-face context. The cognitive interview is conducted using verbal probing techniques such as "think-aloud" to elicit thinking about each question. These probes take a number of forms such as:

- Comprehension probe: "What does the term "dental sealant" mean to you?"
- Paraphrasing: "Can you repeat the question in your own words?"
- Confidence judgment: "How sure are you that your health insurance covers…"
- Recall probe: "How do you know that you went to the doctor 3 times…?"
- Specific probe: "Why do you think that breast cancer is the most serious health problem?"
- 'Back-pocket' probe: "Can you tell me more about that?"

The goal of using these probes is to note apparent problems related to question wording, ordering, and format and then to suggest modifications that address the problems. Best practices suggest doing this as an iterative process consisting of multiple testing rounds.

There are a number of questions that have been raised about the use of cognitive interviewing, some of which have been addressed by research and others that still require future research to be conducted. For example, despite the widespread use of cognitive pretesting to evaluate questionnaires, particularly in government survey labs, it is unclear whether or not independent researchers testing the same questionnaire would reach the same conclusions. Preliminary research has provided promising results about the reliability of the cognitive pretesting findings, but existing research has been limited and incomplete. This indicates a promising and much-needed area for future research on the cognitive pretesting method. A key question needing to be addressed is under what conditions are cognitive interviewing results stable and reliable, and what researchers can do to enhance those conditions.

Additional research is also needed on best practices for designing cognitive pretesting studies themselves. For example, because cognitive interviewing is a qualitative research endeavor, it is often unclear what sample sizes are necessary. Identifying best practices with regard to cognitive interview sample size is important for two reasons, first it is necessary to know how many interviews will be enough to identify a problem and then how many more will be necessary to assess the seriousness or impact of the problem. One recent study has examined this issue and found that, "additional interviews continued to produce observations of new problems, although the rate of new problems per interview decreased" (Blair and Conrad, 2011, p. 654). This finding

needs further study and replication. Developing future research on these questions are important so that researchers can make the most of the resources invested in cognitive interviewing.

It is also important for future research to focus on identifying the utility of cognitive interviewing for mixed-mode surveys and novel administration methods as survey research moves into the future. Much of cognitive interviewing research to date has focused on differences between administration modes because the cognitive issues that appear in self-administered modes are somewhat different from those that are interviewer-administered. Increasingly, the focus has shifted to identifying cognitive issues surrounding web usability and Internet administered surveys, but this is very new and requires significant future research. Other new areas of research have looked at pushing cognitive interviewing itself to different modes such as Skype or other Internet-based approaches to soliciting feedback from participants. For example, research has been done on providing the question probes to respondents to a web survey after each evaluated question using an open text box for them to provide their responses. This practice enables many more cognitive interviews to be performed for the same cost, but it is unclear what is lost in terms of information that an interviewer may have been able to obtain. There is the option of conducting some traditional in-person interviews in tandem with Internet-based approaches to try to maximize the value of both. However, essentially no research has examined this, and it is very necessary in order to adapt cognitive interviewing to the future of survey research.

Finally, more research is needed on applications of cognitive interviewing techniques for addressing issues surrounding cross-cultural comparability within and between surveys. Although cross-cultural differences have been widely recognized by survey researchers, with careful steps taken in sampling, language of administration, and weighting, relatively little has been done with cognitive interviewing to test the differences in cognitive problems that different cultural groups may have with a questionnaire. Further, researchers have not established whether current cognitive interviewing techniques are applicable across cultures, meaning that significant research is needed in this area. Once appropriate cognitive interviewing techniques are identified, they can be applied to ensure that surveys exhibit cross-cultural measurement comparability. A related issue arises from linguistic and translational issues in cross-cultural surveys, which cognitive interviewing should theoretically be able identify. Even basic translations can go very badly if good evaluation and pretesting practices are ignored. In short, cognitive interviewing holds great promise for increasing the ecological validity of survey research in increasingly diverse research contexts, but considerable research is needed to maximize the value of the method.

Areas for future research:

- Identifying under what conditions cognitive interviewing results are stable and reliable
    - What steps can researchers take to enhance those conditions?
- Understanding how many cognitive interviews are necessary to:
    - a) Identify a problem (Number of interviews before Problem X occurs)
    - b) Validate a problem (Of X interviews, problem occurs in at least Y cases)
- Identifying the utility of cognitive interviewing for mixed-mode surveys
- Identifying and testing novel administration methods for cognitive interviews
- Identifying the applicability of cognitive interviewing methods across cultures

- Identifying best practices for using cognitive interviewing to increase cross-cultural comparability

Reference:

Blair, J. and Conrad, F.G. (2011). Sample Size for Cognitive Interview Pretesting. *Public Opinion Quarterly*, *75*(4), 636–658.

Survey Interviewing: Deviations from the Script – Nora Cate Schaeffer

Survey interviews are typically scripted in such a way that the interviewer is intended to follow a particular pre-defined path through the survey, is expected to read questions exactly as they are worded, and is expected to avoid deviations from the survey materials with regard to the content of the questionnaire. However, there are at least two different ways of thinking of the "script" of a survey; the first is the script of the survey questions themselves and the second is the script of the rules of standardization for administering the survey.

Interviewers are the interface between the organization or researcher and the respondent. As such, their behavior during the survey interview is extremely important. Looking at the changes in major research studies that have occurred in the last decade, it is possible to guess the following about future studies: In addition to the sorts of opinion studies or other studies currently being conducted, there will be a class of research studies that will be very complex and demanding for both respondents and interviewers. As the cost of reaching sample members increases, from the researcher's point of view it is economically sensible to ask face-to-face interviewers to do many complex tasks once the interviewers have persuaded the sample member to become a respondent. For these complex interviews to be successful, we need to understand more about how measurement is accomplished within that interaction and how to motivate respondents.

There are many factors that influence the behavior of interviewers. One model of interviewer behavior is an interactional model of the survey response process (Dykema et al., 2013a, 2013b), which is a further development of the model presented in Schaeffer and Dykema (2011b) and can be seen below:

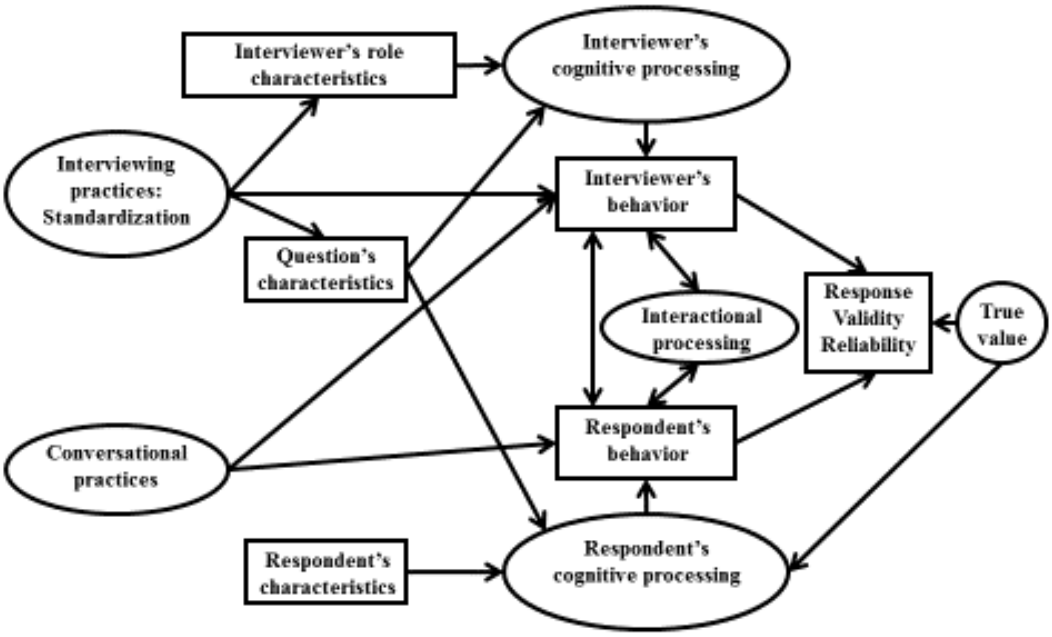


Figure 1. Interactional model of survey response

The model proposes that the main influences on the behavior of interviewers – and their deviations from the script – are:

- Training in the interviewing practices of standardization
- Technology, which has both a direct effect on the interviewer's behavior and an indirect effect through the way technology shapes and limits the characteristics of the survey question
- The characteristics of the survey question (see Schaeffer and Dykema, 2011a), which affects the interviewer's behavior directly as she reads the question and indirectly by the way the question affects the cognitive processing of the respondent and the respondent's subsequent behavior
- The behavior of the respondent, which may require that the interviewer respond in ways that challenge her compliance with standardization
- Interactional and conversational practices, some of which may be made more or less relevant by characteristics of the question or the behavior of the respondent

Technology, whether paper or some electronic technology, presents the script to the interviewer in a way that is often incomplete, so that the interviewer must improvise using the principles of standardization. For example, a paper grid may allow the interviewer to have an overview of the structure of the task and also allow her to enter information that the respondent provides before the interviewer requests it (e.g., the ages of other members of the household). A computer-

assisted personal interviewing (CAPI) instrument, on the other hand, may require that each piece of information be entered on a separate screen, and these constraints may in turn motivate the interviewer to reinforce that standardized order with the respondent.

Interactional and conversational practices are sometimes in tension with standardization. Some of these tensions may have only minor consequences for data quality. For example, interviewers routinely use "okay" both as a receipt for an answer and to announce an upcoming return to the script of the next question. Other tensions may be more consequential. For example, when cognitive assessments or knowledge questions are administered to respondents, the respondent's perception that they are being tested may cause discomfort that leads to laughter, self-deprecating remarks by the respondent, or reassurance by the interviewer (Gathman et al., 2008). These social tensions are not provided for in the rules of standardization, and the improvisations by the interviewer may affect the respondent's willingness to engage in further disclosures or their level of motivation. Interviewing practices and the training that interviewers receive in how to behave in standardized ways also shape the interviewer's behavior. Future research on how conversational practices enter into the interaction between interviewer and respondent will help researchers understand better which behaviors might affect the quality of measurement and how to adapt standardized interviewing practices to changing technology and to maintaining the motivation of the respondent.

One occasion on which interviewers deviate from the script occurs when respondents volunteer a lot of information all at once. This conversational practice presumably is occasioned by the respondent's inference about what the interviewer may ask next. So the respondent may tell the interviewer, "Everybody who lives in this household is white," for example. When that happens, if the interviewer is in a situation where the interview schedule instructs her to ask the respondent the race of each household member, the interviewer must quickly determine how to manage the situation. The interviewer still must follow the rules of standardization, but the interviewer now knows the answers to upcoming questions, or at least what the respondent thinks are the answers. When a respondent gives information to the interviewer "prematurely," the interviewer must balance interactional practices that require that she show she has heard and understood what the respondent just said with the practices of standardization. "Verification" or "confirmation" is a label for interviewing practices that some interviewing shops deploy in this situation; for example, "The next question is 'How old were you on your last birthday?' I think you said you were 65. Is that correct?" In situations like this, the interviewer may not ask all the questions they were supposed to ask, or the interviewer may not follow the rules of standardization because they are trying to manage the information that the respondent has supplied.

There are number of different ways or occasions during which interviewers deviate from the survey script, and a summary list is provided below:

- Not reading the question as worded
- During follow-up behaviors for which there are principles of standardized interviewing but not an actual script, such as:
    - Providing definitions authorized by the instrument or project training
    - Feedback

    - Other follow-up behaviors
- When respondents provide information in variable form such as an event history calendar, timelines, or grids
- Intruding conversational practices, which may happen particularly at sites of tension between conversational practices and standardization

There are surprisingly few studies that have examined these deviations from the survey script using a strong criterion and taking into account that respondents are nested within interviewers, and this is an area that would benefit from future research. The studies that have been conducted seem to indicate that changes in question wording make little difference for reliability (Groves and Magilavy, 1981; Hess et al., 1999). Record-check studies that compared answers to records have found that for most questions examined, substantive changes in question wording had no effect. But there are a few instances in which changes increased or decreased response accuracy (Dykema and Schaeffer, 2005; Dykema et al., 1997). No explanation has been identified for these observed differences in the impact of question reading on accuracy, and this warrants future research (see summary in Schaeffer and Dykema, 2011b).

Behaviors of the interviewer other than question reading are also important. "Probing" is a difficult behavior to identify reliably. Some studies refer to "probing" and others to "follow-up." When probing or follow-up occurs, it is almost always associated with lower quality data, regardless of the adequacy of the interviewer's follow-up or their adherence to standardized practices of follow-up. This is presumably because the interviewer's follow-up was occasioned by the inadequacy of the respondent's answer, an inadequacy that the interviewer might not be able to ameliorate by with her follow-up techniques.

Providing definitions is another form of deviation from interview scripts that has been a source of contention between standardization and conversational interviewing practices. Providing definitions can improve the respondent's understanding of complex concepts when the respondent's situation requires that definition. However, there have not been studies that compare the *ad-lib* method of providing definitions with other methods of providing definitions (many of which are discussed by Schober and Conrad (1997) in their paper on this topic). So, it would be helpful for future research on interviewing practices to expand the comparison, not just to providing *ad-lib* definitions versus no definitions, but examining other ways of providing definitions and doing it in a study design that allowed both variable errors and bias to be assessed simultaneously. There is also a need for the right statistical design and analyses to be applied in this sort of research; in particular, future research should ensure that there are sufficient numbers of interviewers and respondents per interviewer. Analyses should model the structure of the data including the hierarchical structure of respondents nested within interviewers.

Deviations can also be thought of as initiated by either respondent behavior or interviewer behavior. In the case of respondents, research suggests that these classes of behavior are associated with tensions in standardization (see Schaeffer and Dykema, 2011b; Schaeffer and Maynard, 2008):

- Informative contributions by respondent
    - Relevant information in place of properly formatted answer

        - Substantive ("just reading glasses")
        - Synonyms for response categories ("probably")
        - Uncertainty markers ("never thought about it")
    - Information accompanying properly formatted answers
        - Information beyond that requested by the initial question but sought by subsequent questions
        - Qualifiers and mitigators
        - Considerations
- Interruptions

These behaviors of respondents can occur because of state uncertainty (meaning that the respondent is unsure of their answer) or because of task uncertainty (the respondent is unsure how to fit their answer into the structure of the task) (Schaeffer and Thomson, 1992). Interviewers may depart from the rules of standardization during different types of actions:

- Posing questions
    - Offering definitions or response categories in questions where they were not scripted
    - Tailoring questions in a series or battery of questions to reduce repetition
- During follow-up actions that confirm or code respondents' answers
    - Tuning – working to get a more precise response from the respondent by only repeating response categories that appear to be in the vicinity of the respondent's answer
    - Verification or confirmation of an answer that was provided in response to a previous question
- Responding to an answer that is not adequate
    - Providing or applying definitions
    - Reducing or simplifying the task
    - Asking follow-up questions that target an ambiguity
    - Repeating questions or response categories
- Giving feedback
    - Receipt answers
        - "Okay"
        - Confirm or repeat the respondent's answer
    - Announce a return to agenda
    - Reinforce and motivate ("Thank you. That's the kind of information we are looking for.")

Some topics or tasks in the survey instrument are probably more common sites for departures from standardization. These include complex topics such as household listings or rosters to determine the structure of a household and event history calendars or complex tasks such as physical measurements, cognitive assessments, and obtaining permission for records linkage. These are just some examples of common areas for which the quality of measurement could be improved by attention to the design of the instrument and the development of appropriate interviewing practices.

Interviewing practices are a complex array of intersecting, and occasionally colliding, demands that interviewers must navigate. It is easy for researchers to focus on ways that interviewers deviate from prescribed behaviors and condemn the negative influence of these behaviors on data quality, but all too often the role of the instrument is ignored. Observers may assume that there is a good reason for every protocol and approach in a survey instrument, and it is the duty of the interviewer to "simply" follow the script and collect the data. In reality, as exhibited in the discussion of rosters, no instrument can provide a complete script, and some elements of the instrument design may be frustrating for both respondents and interviewers. These frustrating aspects of the interview may have neutral or negative (if they increase interviewer variability) effects on data quality. On rare occasions, features of the interview that the respondent finds frustrating may motivate the respondent to break-off and not complete the interview. Although such breakoffs are not common in interviewer-administered interviews, motivation to participate in subsequent interviews could suffer. Any negative consequences of the experience of the interview for respondents add to the complications interviewers face in responding to demands that they achieve high response rates.

These considerations imply that researchers bear some responsibility for fielding instruments that minimize the occasions on which interviewers deviate from good interviewing practices. Interviewing is a constrained and specialized interaction because of the needs of measurement, but it is still an interaction between people, and instrument designers need to bear this in mind. By observing interviewers and respondents in action, researchers can see the problems that interviewers and respondents face and the ways that they solve them; these observations can and should be used to inform instrument design and interviewer training. Lastly, changing and novel interviewing technologies and the varieties of types of information being collected by surveys may require innovations in interviewing practices. Survey researchers should remain keenly aware of the demands that these new challenges place on interviewers and the possible consequences for the quality of measurement.

Studies of interviewing practices require designs that can allow us to draw conclusions because they preserve features of large-scale production surveys, include and document the methods for training and monitoring interviewers, include manipulation checks, and assess reliability or validity or both.

Areas for future research:

- How the interaction between the interviewer and respondent affects both the motivation of the respondent and the quality of the resulting measurement in a way that considers interviewer effects.
- How conversational practices enter into the interaction between interviewer and respondent, their impact on the motivation of the respondent, and the quality of resulting measurement.
- How the interviewing practices that the interviewer uses to manage information that the respondent supplies before it is requested affects the motivation of the respondent and the quality of measurement, and what practices interviewers should use to deal with this information.
- Under what conditions do changes that the interviewer makes in the wording of a question when it is originally read lead to an increase or decrease in the accuracy of responses?

- How to improve practices that interviewers use to follow up answers that express uncertainty so that the quality of the resulting data is improved.
- How effective are different methods for providing respondents with definitions for complex target objects considering that the assessment must
    - Consider both variable errors and bias simultaneously
    - Include methods suitable for long production surveys
- Re-assessing principles of questionnaire design to find methods that reduce the burden on both interviewers and respondents
- Building "smarter" survey systems that integrate and display information previously recorded and allow interviewers to enter answers in the order that the respondent provides them so that interviewers do not need to ask for redundant information
- How real interviewers and real respondents interact with the survey technology and questionnaires in the real world to devise improved question designs and rules for interviewing.
- What interviewing practices do we need for complex interviews of the future that include such complex tasks as physical measurement, cognitive assessments, and so forth.

References and Further Reading:

Belli, R.F. and Lepkowski, J.M. (1996). Behavior of Survey Actors and the Accuracy of Response. In: Warneke, R.B. (Ed.), *Health Survey Research Methods Conference Proceedings, DHHS Publication, No. (PHS) 96-1013,* (pp. 69-74). Hyattsville, MD: Department of Health and Human Services, Public Health Service, Centers for Disease Control and Prevention, National Center for Health Statistics.

Conrad, F.G. and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. *Public Opinion Quarterly, 64,* 1-28.

Dijkstra, W. (1987). Interviewing Style and Respondent Behavior: An Experimental Study of the Survey Interview. *Sociological Methods and Research, 16,* 309-334.

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study, In: Lyberg, L, Biemer, P., Collins, M., de Leeuw, E.D., Dippo, C., Schwarz, N., and Trewin, D. (Eds.), *Survey Measurement and Process Quality,* (pp. 287-310). New York, NY: Wiley-Interscience.

Dykema, J. and Schaeffer, N.C. (2005). An Investigation of the Impact of Departures from Standardized Interviewing on Response Errors in Self-Reports about Child Support and other Family-Related Variables. Presented at the Annual Meeting of the American Association for Public Opinion Research Annual Meeting, Miami Beach, Florida.

Dykema, J., Schaeffer, N.C., and Garbarski, D. (2013a). Associations between Interactional Indicators of Problematic Questions and Systems for Coding Question Characteristics. Paper presented at the annual meeting of the American Association for Public Opinion Research, Boston, MA.

Dykema, J., Schaeffer, N.C., Garbarski, D., Nordheim, R., and Cyffka, K. (2013b). Effects of Question, Respondent, and Interviewer Characteristics on Interactional Indicators of Respondent and Interviewer Processing of Health-Related Questions. Paper presented at the Interviewer-Respondent Interaction Workshop, Boston, MA.

Fuchs, M. (2000). Screen Design and Question Order in a CAI Instrument: Results from a Usability Field Experiment. *Survey Methodology, 26*, 199-207.

Fuchs, M. (2002). The Impact of Technology on Interaction in Computer-Assisted Interviews, In: Maynard, D.W., Houtkoop-Steenstra, H., van der Zouwen, J., and Schaeffer, N.C. (Eds.), *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview,* (pp. 471-491). New York, NY: John Wiley & Sons, Inc.

Fuchs, M., Couper, M., and Hansen, S.E. (2000). Technology Effects: Do CAPI or PAPI Interviews Take Longer? *Journal of Official Statistics 16*(3), 273-86.

Gathman, E., Cabell, H., Maynard, D.W., and Schaeffer, N.C. (2008). The Respondents are all Above Average: Compliment Sequences in a Survey Interview. *Research on Language and Social Interaction, 41*, 271-301.

Groves, R.M. and Magilavy, L.J. (1981). Increasing Response Rates to Telephone Surveys: A Door in the Face for Foot in the Door. *Public Opinion Quarterly, 45,* 346-358.

Hak, A. (2002). How Interviewers Make Coding Decisions. In: Maynard, D.W., Houtkoop-Steenstra, H., van der Zouwen, J., and Schaeffer, N.C. (Eds.), *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview,* (pp. 449-470). New York, NY: John Wiley & Sons, Inc.

Hess, J., Singer, E., and Bushery, J. (1999). Predicting Test-Retest Reliability from Behavior Coding. *International Journal of Public Opinion Research*, *11*(4), 346-360.

Mangione, T.W., Fowler, F.J., and Louis, T.A. (1992). Question Characteristics and Interviewer Effects. *Journal of Official Statistics, 8*, 293-307.

Moore, R.J. and Maynard, D.W. (2002). Achieving Understanding in the Standardized Survey Interview: Repair Sequences. In: Maynard, D.W., Houtkoop-Steenstra, H., van der Zouwen, J., and Schaeffer, N.C. (Eds.), *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview,* (pp. 281-312). New York, NY: John Wiley & Sons, Inc.

Schaeffer, N.C. and Dykema, J. (2011a). Questions for Surveys: Current Trends and Future Directions. *Public Opinion Quarterly, 75*(5), 909-961.

Schaeffer, N.C. and Dykema, J. (2011b). Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions In: Madans, J., Miller, K.,

Maitland, A., and Willis, G. (Eds.), *Question Evaluation Methods: Contributing to the Science of Data Quality,* (pp. 23-39). Hoboken, NJ: John Wiley & Sons, Inc.

Schaeffer, N.C. and Maynard, D.W. (2008). The Contemporary Standardized Survey Interview for Social Research, In: Conrad, F.G. and Schober, M. F. (Eds.), *Envisioning the Survey Interview of the Future,* (pp. 31-57). Hoboken, NJ: John Wiley & Sons, Inc.

Schaeffer, N.C. and Thomson, E. (1992). The Discovery of Grounded Uncertainty: Developing Standardized Questions about Strength of Fertility Motivation. In: Marsden, P.V. (Ed.), *Sociological Methodology 1992,* Vol. 22, (pp. 37-82). Oxford: Basil Blackwell.

Schnell, R. and Kreuter, F. (2005). Separating Interviewer and Sampling-Point Effects. *Journal of Official Statistics, 21*. 389-410.

Schober, M.F. and Conrad, F.G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? *Public Opinion Quarterly, 61*, 576-602.

Suchman, L. and Jordan, B. (1990). Interactional Troubles in Face-to-Face Survey Interviews. *Journal of the American Statistical Association, 85*, 232-253.

Challenges and Opportunities in Open-Ended Coding – Arthur Lupia

In surveys, responses to open-ended question are often released to the public in coded, categorized forms. The typical rationale for converting open-ended responses to coded categories is to protect respondent privacy. Specifically, survey participants sometimes express their responses in ways so unique that others could use that information to identify them. Many leading surveys have concluded that the public release of such information is not worth the risk to respondent privacy and produce codes instead. Many survey researchers, moreover, like the apparent convenience associated with using coded data rather than having to work on their own to translate words into statistically analyzable elements.

Unfortunately, methodological approaches to open-ended coding in survey research have been anything but consistent and credible. The lack of widely accepted best practices might be partially responsible for the decline in the use of the open-ended response format. This is unfortunate because poorly defined methods and best practices should not prevent the use of open-ended responses that can often provide a uniquely rich source of data to survey researchers.

The fundamental question that researchers seeking to code open-ended data must address is, "What is the correct inference for a user to draw from a coded open-ended response to a survey question?" The answer to that question is going to depend on what question the survey asked, what the respondent says, and then very importantly, on processing decisions that are made after the interview is conducted about converting respondent answers into numbers. A key point is that researchers creating and using these data need to pay greater attention to possible problems caused by inattention to these details.

For this discussion, examples will be drawn heavily from the 2008 American National Elections Study (ANES), so it is worth outlining the key features relating to open-ended questions on the ANES. There are four general types of questions that the ANES asked in the open-ended format: 1) "most important problem," 2) candidate "likes-dislikes," 3) party "likes-dislikes," and 4) political knowledge.

The "political knowledge" questions are noteworthy because they are one of the most frequently used variables that the ANES produces. These questions solicit answers to fact-based quiz questions about politics such as:

> "Now we have a set of questions concerning various public figures. We want to see how much information about them gets out to the public from television, newspapers, and the like…. What about … William Rehnquist – What job or political office does he NOW hold?"

This question is administered in an open-ended format, allowing respondents to answer in their own words. In most cases, the participant's complete response has been recorded verbatim. Later, all answers have been coded as either "correct" or "incorrect." For decades, the publicly available versions of these variables have been treated as valid measures of respondents' knowledge of the question. However, questions have been raised about the accuracy and relevance of this data.

Generally, ANES users expect the study investigators to convert open-ended responses into numeric format coded as correct or incorrect. A critical point about this process is that the users base their inferences on beliefs about what each of these numbers means. Many users believe that open-ended coding is easy to do, that it generates valid measures, and that it's performed well by survey organizations. The evidence tells a different story. For the 30-years that the ANES has been asking open-ended recall questions, there is little to no record of users asking critical questions such as, "How did you make decisions about what answers were correct or incorrect?" There is a similarly sparse record of requests for reliability statistics, and yet, the questions are widely used under the assumption that the data are valid and reliable.

Moreover, after examining the open-ended question data collected in the 2008 ANES under more intense scrutiny, researchers found that the reality of the state of open-ended coding is much worse than many users assumed. Coding practices for other major surveys were examined and found to be similarly disappointing. So it is important to evaluate where things have gone wrong.

Consider, for example, problems and controversies associated with the ANES question about the position that William Rehnquist held. Many studies have examined this and similar items and drawn such conclusions as:

> "Close to a third of Americans can be categorized as "know-nothings" who are almost completely ignorant of relevant political information, which is not, by any means, to suggest that the other two-thirds are well informed (Bennett, 2006).

> "The verdict is stunningly, depressingly clear: most people know very little about politics." (Luskin 2002)

Indeed, a study by Gibson and Caldiera (2009) found that only 12% of respondents provided a "correct" response to the Rehnquist question. At first glance, this statistic appears to support the conclusions. However, a closer examination of the coding of the responses to the question found that, for the 2004 ANES, responses were marked as "correct" only if respondents specifically identified that Rehnquist was "Chief Justice" and on the "Supreme Court" – meaning that the additional 30 percent of respondents that identified him as a Supreme Court justice were marked "incorrect" due to not specifying "Chief Justice". When Gibson and Caldiera (2009) asked respondents to state whether William Rehnquist, Lewis F. Powell, or Byron R. White was Chief Justice, 71% correctly identified Rehnquist. Similarly, on the 2000 ANES, 400 of the 1,555 respondents said that Rehnquist was a judge or said that he was on the Supreme Court but were coded as having answered incorrectly. Other "incorrect" answers from the 2000 ANES included:

- "Supreme Court justice. The main one."
- "He's the senior judge on the Supreme Court."
- "He is the Supreme Court justice in charge."
- "He's the head of the Supreme Court."
- "He's top man in the Supreme Court."
- "Supreme Court justice, head."
- "Supreme Court justice. The head guy."
- "Head of Supreme Court."
- "Supreme Court justice head honcho."

Along similar lines, another political knowledge question asked "…Tony Blair - what job or political office does he NOW hold?" For this item, there was a serious error in the coding instructions. ANES coders were told that in order to be marked as "correct", "the response must be specifically to 'Great Britain' or 'England' – United Kingdom is *NOT* acceptable (Blair was not the head of Ireland), nor was reference to any other political/geographic unit (e.g. British Isles, Europe, etc.) correct." In fact, Blair was "Prime Minister of the United Kingdom". While many scholars and writers used this variable to characterize public ignorance, there is no record that any such user questioned how it was produced prior to subsequent ANES PIs discovering this error.

With this example in hand we return to questions about how do these kinds of things happen and what can be done to improve open-ended coding practice. The answer to the first question is that researchers don't really know due to poor recordkeeping prior to 2008. Typically for the ANES, interviewers would transcribe the respondent answers, and staff would then implement a coding scheme, often with cases coded by a single person. However, there are few or no records of instructions given to staff regarding coding and there is no documentation of coding reliability analyses. We later discovered that such recordkeeping inadequacies are not particular to the ANES and are common across major surveys.

In response to these discoveries, the ANES moved to make redacted transcripts available whenever possible, conducted a conference to discover and develop best practices, and formed

expert committees to develop mutually exclusive and collectively exhaustive coding schemes. The key outcomes included setting clear definitions for which responses would be coded as "correct" and "incorrect" or "partial knowledge".

The expert discussion about "partial knowledge" was particularly interesting. Recall that the question begins, "What's the job or political office" that a particular person holds. For example, when this question was asked about Dick Cheney in 2004, a lot of people provided the correct response of "Vice-President". Others, however, would say "anti-Christ" or "chief puppeteer". How should these responses be coded? Some experts wanted to count such responses as representing partial knowledge. A breakthrough occurred when the experts returned to the question wording and determined to focus the coding on whether or not a correct answer was given to the question that the ANES actually asked. So if a respondent says that Dick Cheney shot his friend in a hunting accident, they have decided not to answer the question. They may be providing knowledge about Dick Cheney, but still not answering the question as it was asked. So the whole concept of "partial knowledge" was changed by asking, "Is it partial knowledge with respect to this question?"

The ANES's new coding framework for open-ended questions about political office place emphasis on the following factors:

- Did the respondent say something about the political office?
- Did they identify any part of the title of this person's political office correctly?
- Some people have multiple offices; for example, the Vice-President of the United States is also President of the Senate. The person who is the Speaker of the House is also a congressperson.

The first factor on which the new coding scheme focuses is "did the respondent say anything correct or partially correct about the political office?" Since the question asks about the person's job, a correct answer to the question that was asked would also constitute descriptions of what this person does, of what their job is, so things like making legislation or organizing a political party would also constitute a correct or partially correct response. So for each of the questions the ANES asked, a long list of jobs that a particular person has was identified. So, unlike before where the ANES produced a single "correct" or "incorrect" code for each question, now they have a political office code: "Does the respondent give a correct answer regarding the political office?" The ANES also has a code to indicate if the respondent provided a complete description of a job or an incomplete description. And then finally there is an "other" code, and "other" is anything that the person says that is not pertaining to the job or political office of that person. With respect to the "other" responses, the ANES coding system does not make any judgments about them. In sum, the codes reflect the question, "Did the respondent name the political office, did they name the job, or have they said anything else?"

In terms of procedural transparency, the ANES has decided that it is important for users to be able to see how the code frame was implemented so they can figure out whether there is something about the coding practices that skew the numbers. Written documentation of all decisions and written documentation of all conversations that the ANES had with coding vendors, particularly regarding instructions that confused coders, are now available on the ANES website.

The goal of making such information available is for people who are using the data, or who are developing their own coding schemes, to be able see what the ANES did. With such information in hand, scholars who obtain different results than the ANES have better ways of determining "Why?"

Several attributes of the new ANES coding scheme provide some insight into what may be best practices for other surveys or researchers to consider when coding open-ended responses. First, the scheme is theoretically defensible; second, it has demonstrated high inter-coder reliability; third, it is mutually exclusive and collectively exhaustive; and fourth, it is accessible to other scholars who may want to use public data to compare other code frames. More general steps that have been taken include:

- Increased documentation at all stages
- Evaluation at many stages
- Increased procedural transparency
- High inter-coder reliability

In summary, if researchers believe that the scientific credibility of survey-based research depends on transparency and replicability of analysis, then it is imperative that data be manufactured and distributed in ways that facilitate transparency and replicability. The kinds of practices needed to help survey producers and analysts make more effective decisions about how to code open-ended data and interpret the resulting variables are not what they could be. Future research should identify ways to improve practice (and data quality) in ways that need not require extra resources to implement. Doing so should give a wide range of scholars the direction and motivation needed to increase documentation of past coding schemes and engage in more systematic thinking about how best to develop future coding algorithms.

Areas for future research:

- Developing effective approaches to increase coding transparency
- Identifying ways to develop shared coding schemes and practices between survey organizations that are asking the same questions
- Specifying best practices regarding the production and dissemination of documentation

References:

Bennett, S.E. (2006). Democratic competence, before Converse and after. *Critical Review, 18*(1-3), 105-141.

Gibson, J.L. and Caldeira, G.A. (2009). Knowing the Supreme Court, A Reconsideration of Public Ignorance of the High Court. *The Journal of Politics, 71*(2), 429-441.

Luskin, R.C. (2002). From Denial to Extenuation (and Finally Beyond): Political Sophistication and Citizen Performance. In: Kuklinski, J.H. (Ed.), *Thinking about Political Psychology,* (pp. 281-305). New York, NY: Cambridge University Press.

What Human Language Technology can do for you (and vice versa) – Mark Liberman

Human Language Technology (HLT) is a broad class of approaches to taking human language in its various written and spoken forms and making its content into data that is accessible to computational analysis. The term HLT comes from work sponsored over the past three decades by the Defense Advanced Research Projects Agency (DARPA) in its Speech and Language program. HLT is a more general term for techniques that are described by phrases such as Natural Language Processing (NLP), Computational Linguistics, Speech technology, etc.

When the input is text, HLT refers to tasks that include:

- Document retrieval ("Googling")
- Document classification
- Document understanding
- Information extraction from text
- Summarization
- Question answering
- Sentiment analysis (opinion mining)
- Machine translation

When the input is speech, HLT tasks include:

- Speech recognition ("speech to text")
- Speech synthesis ("text to speech")
- Speaker identification/classification
- Language identification
- Diarization ("who spoke when")
- Spoken document retrieval
- Information extraction from speech
- Question answering
- Human/computer interaction via speech
- Speech-to-speech translation

As a result of the past decades of research, HLT methods are now good enough that survey researchers are increasingly looking to them as a means of quickly coding open-ended responses to survey questions without needing to hire and train human coders. HLT researchers have been aware of the potential to apply these methods to open-ended survey questions for over 15 years, yet they have not seen widespread use in survey research.

There are a number of potential reasons for the slow adoption of HLT in survey research. First, HLT applications might not work well enough to replace human coders in this context, because the error rate that researchers are willing to accept is relatively low and the variability in results between different HLT methods can be quite high. Second, early attempts at using HLT for survey applications were seen as unsuccessful, so that many people may write off the approach even though the technology has improved considerably. Third, generic out-of-the-box HLT

solutions will not always work to solve particular survey research problems, and there has not been enough demand for survey-specific research to generate methods targeted at this application. This is further complicated by the fact that the process of figuring out whether or not any given pre-existing approach will work in a particular survey context may be somewhat protracted and difficult.

One reason that HLT applications in the survey context can be particularly difficult is the fact that human annotation of text, in the absence of careful definition and training, is extremely inconsistent. If you give annotators a few examples, or a simple definition, and turn them loose, the resulting level of inter-annotator agreement is generally quite poor. This is true even in the identification of apparently well-defined entities, which are things like people, places, organizations, or, in a biomedical domain, genes, gene products, proteins, disease states, organisms, and so on. It is even harder to annotate relationships among entities in a consistent way.

For the sake of understanding the complexity of the annotation task in identifying entities, consider an example: If you were to take a large group of Ph.D. political scientists, give them scientific papers that are about areas in their specialization, and ask them if they can determine when politicians are mentioned in those papers, they'll look at you like you're an idiot because the task seems so simple. However, if you were to take two of them, put them in separate rooms, ask them to do the same task, and then look at how well they agree on the output, you would be very fortunate if they agree 50 percent of the time. It gets even worse for what are called "normalized entities." That is, instead of simply asking, "Is this referring to a politician?", you want to know which individual is referred to.

Here are some of the issues that will come up: Which categories count as "politicians" (judges, attorneys general, political pundits, appointed officials, protest leaders, etc.)? Do references to groups count (e.g. "the members of the Warren court")? What about names used as modifiers ("Stalinist techniques")? What about specific but as yet unknown individuals ("the Republican nominee for president in 2016")?

These problems arise because human generalization from examples to principles is variable, and human application of principles to new examples is equally variable. As a result, the "gold standard" is not naturally very golden. The resulting learning metrics are noisy, and an F-score (the harmonic mean of precision and recall) of 0.3 or 0.5 is not a very attractive goal. If researchers tell people that they can write a program that will agree with their intuitions 30 percent of the time, they're not very impressed, even if their intuitions and their neighbor's intuitions agree only 30 percent of the time.

For research on information extraction from text, HLT researchers have developed an iterative approach that can achieve relatively high rates of agreement among human annotators. This process is analogous to the development of common law: a relatively simple statute says what the rules are, and then a long list of particular cases provide precedents that can be applied to new instances. (Of course, survey researchers face analogous problems in deciding how to classify the answers to open-ended questions.)

The resulting guidelines for HLT annotation tasks can sometimes be hundreds of pages long. These approaches are slow to develop and hard to learn, but in the end they can produce inter-annotator agreement rates of 90% or better. And with adequate amounts of training data of this type, there are standard ways to create automated systems whose agreement with human annotation is nearly as good.

So, while HLT holds great promise for automating the coding of open responses to survey questions, there is still a considerable amount of research that is needed before it will see widespread application in practice.

This is an important area for ongoing research for at least two reasons. First, it will reduce the costs of analyzing open responses to surveys and enable the open response format to be more widely used. As was mentioned in the question design section above, this format is an important source of high quality survey data. Second, a considerable amount of archival open response data could be analyzed in a more consistent manner than the current patchwork approach of having different teams of human coders conduct coding at different points in time.

While it seems like a daunting task to reconcile the current state of HLT with respect to the needs of survey research, it is important to continue pushing research forward into unknown and untested domains. Allan McCutcheon made a key point to this effect during the Q&A for this talk when he stated:

> *"We can't plan for the technology of today. We've got to start thinking about the technology of tomorrow. Right? And the technology – I mean if you told people ten years ago that you'd be talking **to** your cell phone, they would have said, 'Well, yeah, to my mother.' Right? But, 'No, no, no, you'll be talking **TO** your cell phone.' Right? They'd look at you as if you just had your head screwed on wrong. But today, people are doing it, and they're saying, 'Well, it doesn't do it perfectly. It doesn't understand me as well as my mother does.' Wait for ten years."*

Areas for future research:

- Developing survey-specific HLT applications for coding open response data
    - Identifying the unique features of survey data that will need to be addressed by the statistical algorithms in HLT
    - Identifying similar projects in other research domains that could benefit from the development of HLT in the survey domain to maximize intellectual cross-fertilization and share development costs
- Identifying additional external sources of text that can be analyzed and coded by HLT and linked with survey data (e.g. government or medical records)
- Applications of sentiment analysis to survey responses (including even yes/no responses).
- Identifying ways for HLT and survey research to simultaneously break new ground via collaborative research projects that will benefit both fields

Confidentiality and Anonymity – Roger Tourangeau

An important concern for all researchers collecting survey data is respondent confidentiality. Even surveys that do not collect sensitive data need to be concerned about confidentiality, since it affects willingness to participate in surveys and thus may influence response rates and nonresponse bias.

Privacy concerns can be thought of as involving a respondent's unwillingness to reveal information at all; respondent feelings or beliefs that "it's none of your business" in response to a request or potential request for information reflect concerns about privacy. Even respondents who are willing to divulge information to the researchers may not want that information shared with anyone else; concerns about the latter are confidentiality concerns. And there are at least two different classes of third parties that respondents might be concerned about: It could be that somebody in the respondent's household might overhear what he or she said during an interview and might learn something that the respondent would rather they didn't know. Or it could be that some third party outside the household will get hold of the data, maybe a criminal or another federal agency.

Many survey interviews are not done in private. Some research on the American National Election Studies (ANES; Anderson et al., 1988) suggests that up to 40-50 percent of the interviews are done with some other household member present. Other evidence comes from the World Mental Health Surveys (Mneimneh, unpublished dissertation), which are conducted in many countries around the world. The findings indicate that there is wide variability across participating countries in interview privacy conditions. For example, in Japan, about 13 percent of the interviews were done with somebody else present, whereas in India 70 percent are done with another person present. These findings are more than a little troubling because a lack of privacy may influence respondent willingness to report certain attitudes or behaviors.

There are three major groups of factors that affect whether someone beside the interviewer or respondent is likely to be present during a survey interview. The first set involves the household's characteristics – its size, whether the respondent is married, and, if so, whether the spouse is employed. Second are cultural norms regarding privacy. In some countries, respecting people's privacy is a value; in other countries (e.g., those with collectivist cultures), sharing with other people is more important, and privacy is not an important value. A third set of variables involves the amount of effort that interviewers make to provide a private interview context. There seems to be much variation across interviewers in how much they understand that the interview is supposed to be done in private and in how much effort they make to provide those conditions for the respondent.

The consequences of a lack of privacy also vary depending on several factors. First, it depends on whether the other person present during the interview already knows the information being requested of the respondent and, if not, whether there are likely to be repercussions if he or she finds out the respondent's answer. Perhaps as a result, there are typically lower levels of reporting of sensitive behaviors when the respondent's parents are present but fewer effects when the respondent's spouse is present. More generally, there is evidence of increased social desirability bias when interviews are not done in private.

One approach to dealing with the issues surrounding privacy and confidentiality concerns is to offer anonymity to the respondents. This is often hard to do convincingly, especially in face-to-face surveys where the interviewer clearly knows the household address and may also elicit the respondent's signature on a consent form. However, some studies attempt to collect data anonymously despite these challenges. Monitoring the Future, for example, is a study of high school seniors about drug use, and its questionnaires are sent to schools, where they are distributed in classrooms. There is no identifying information on the questionnaires. Even in the cases where this can be accomplished, there are concerns about the potentially negative effects of anonymity on respondent accountability and data quality (Lelkes et al., 2012).

The three items discussed thus far, privacy, confidentiality, and anonymity, are typically discussed in the context of another issue, which is asking sensitive questions. There are at least three distinct meanings for "sensitive question":

- Intrusiveness: The question is inherently offensive (thus, it is the question rather than the answer that is sensitive);
- The risking of disclosure to third parties (various types of third parties), including:
  - Other family members or persons,
  - Other agencies, or
  - Analysts or hackers (disclosure avoidance methods designed to reduce likelihood that this will happen); and
- Social desirability (Socially approved vs. socially disapproved answers)

Social desirability is the focus of much of the attention given to confidentiality, privacy, and anonymity in survey research. In a classic description of the problem, Sudman and Bradburn (1974) say:

> "*Some questions call for the respondent to provide information on topics that have highly desirable answers ... If the respondent has a socially undesirable attitude or if he has engaged in socially undesirable behavior, he may … desire to appear to the interviewer to be in the socially desirable category. It is frequently assumed that most respondents resolve this conflict in favor of biasing their answer in the direction of social desirability*" (pp. 9-10).

There are three primary concerns about the consequences of question sensitivity. The first is unit nonresponse, meaning that people may fail to participate at all if they think they will be asked sensitive questions. The second is missing data, meaning that some respondents may skip offensive or embarrassing questions. The third and perhaps greatest concern is reporting errors. This refers to respondents overreporting desirable attitudes or behaviors and underreporting undesirable ones, in either case, providing false information to the researcher.

Fortunately, researchers have come up with a number of techniques for addressing these concerns about reporting errors. First, self-administration seems to help because respondents no longer need to worry about self-presentation to the interviewer. Some surveys are primarily interviewer-administered, but the sensitive questions are asked in a separate self-administered

section. Second, open-ended responses have been demonstrated to provide better data than closed items. Finally, the randomized response technique and bogus pipeline approaches (both described in more detail below) have shown promise for reducing inaccuracy in reporting. Two key early papers on these topics are Locander et al. (1976) and Blair et al. (1977).

A popular approach to minimizing social desirability bias has been the randomized response technique (RRT). This involves estimating the prevalence of some characteristic without knowing what question any specific respondent received. For example, statement A might be, "I am for legalized abortion on demand," and statement B is, "I'm against legalized abortion on demand." Respondents get one or the other of these items with some known probability. The most common randomization approach is a coin flip. RRT often seems to work, in that researchers get a higher estimate of various sensitive characteristics under RRT than from a direct question. However, no production survey uses this method, because it is difficult to implement in the field and because it increases the variance of the estimates. Because of the impracticality for real application to large production surveys, it is recommended that no further funding be allocated to research on the randomized response technique.

Other clever methods have also been developed, including the item count technique (ICT) and the bogus pipeline, but it's not clear whether they add much in terms of being widely applicable as approaches to reducing measurement error due to sensitive questions. These approaches (RRT, ICT, bogus pipeline, etc.) may have promise for certain very specific applications, but none of them is sufficient to address the real scope of the problem posed by sensitive questions. RRT and ICT have the additional drawback that they do not produce individual estimates of those variables, and only aggregate statistics can be formed, which reduces their utility further.

Researchers often worry about privacy and confidentiality, but many surveys are still done in the presence of other people; this implies that perhaps more training is needed to impress on interviewers that privacy is an important condition to achieve for conducting interviews. Second, measurement error can be a very large problem on surveys that ask sensitive questions, often swamping other sources of error, at least at high levels of aggregation. This suggests that more studies should be done to identify the specific, most worrisome sources of error for each survey, because there is wide variability in how different types of error affect different surveys. Certainly, surveys in general should continue focusing resources on issues like sampling bias, but for some surveys it is likely that measurement error due to sensitive questions is a larger component of the total survey error. Thus, reducing measurement errors should be a goal to which more resources are devoted. Third, self-administration seems to help reduce reporting error, but it is not sufficient to eliminate it; there is still considerable room for improvement.

In terms of future research, we need to devise new methods for collecting accurate information on sensitive topic. And, in thinking a little bit more about recommendations to the National Science Foundation, the research that, in my view, ought to be funded falls under three headings: causes, consequences, and fixes.

Under the "causes" heading: First, most researchers focus on "lying" by respondents, but that term may be too strong for what's really going on. My colleagues and I have used the phrase "motivated misreporting", but we just don't really understand the processes that lead to

misreporting very well. What is it that people are thinking, and what are they doing? These are key areas for future research. It may be a semiconscious process that influences these misreports. It's possible that people are so adept at dodging embarrassing questions in everyday life, that they do it unthinkingly in surveys. It may be a kind of conversational skill carried over into the interview setting. Many researchers seem to think that once they invoke "social desirability," they're done, and that that's the explanation for the measurement errors. The second area on the causal side where more work is needed is on what topics people regard as embarrassing. The presumption is that these potentially embarrassing topics are the ones they'd lie about, but that may not be true either. Future research is needed to develop a firm understanding of what topics are really sensitive and for whom. We need research to understand both the processes by which people modify their answers and the determinants of sensitivity that lead them to do this.

In terms of "consequences": First, more studies are needed to evaluate the relative magnitude of the different sources of survey error. These studies are tough to do, but they will help researchers to avoid simply shooting in the dark. It would be good to have a body of a hundred studies, not just two or three, that look at this issue. Then researchers could say, "These hundred studies have looked at the relative magnitudes of the different sources of error, and they suggest that we really ought to be worrying about X, at least when the topic is Y." Second, researchers need to be cleverer about record check studies and other forms of validating information provided by respondents.

Regarding "fixes" for these problems, it is very unclear what the right strategies are given the current state of research. Researchers have spent a lot of time over the years on things like randomized response techniques. But moving forward, further variations on RRT are not part of the solution. This research needs to go in some new directions and break away from techniques that are either ineffective or not widely applicable.

Thus, the key areas for future research are:

- Identifying best practices for ensuring that surveys are conducted in private settings without other people present
- Mapping the cognitive processes underlying inaccurate answers to sensitive questions
- Evaluating the relative magnitude of different sources of survey error for individual surveys
- Developing more and better approaches to validating respondent answers to get a better sense for actual levels of misreporting on particular items
- Finding new practical methods for collecting sensitive information

References:

Anderson, B. A., Silver, B. D., and Abrahamson, P. (1988). The Effects of the Race of the Interviewer on Measures of Electoral Participation by Blacks in SRC National Election Studies, *Public Opinion Quarterly, 52*, 53-88.

Blair, E., Sudman, S., Bradburn, N.M., and Stocking, C. (1977). How to Ask Questions about Drinking and Sex: Response Effects in Measuring Consumer Behavior. *Journal of Marketing Research, 14*(3), 316-321.

Lelkes, Y., Krosnick, J.A., Marx, D.M., Judd, C.M., and Park, B. (2012). Complete Anonymity Compromises the Accuracy of Self-Reports. *Journal of Experimental Social Psychology, 48*(6), 1291–1299.

Locander, W.B., Sudman, S., and Bradburn, N.M. (1976). An Investigation of Interview Method, Threat, and Response Distortion. *Journal of the American Statistical Association, 71*, 269–275.

Mneimneh, Z. (2012). Interview Privacy and Social Conformity Effects on Socially Desirable Reporting Behavior: Importance of Cultural, Individual, Question, Design and Implementation Factors. Doctoral Dissertation, University of Michigan-Ann Arbor Program in Survey Methodology. Available at deepblue.lib.umich.edu.

Sudman, S. and Bradburn, N.M. (1974). *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine.

Respondent Attrition vs. Data Attrition and Their Reduction - Randall J. Olsen

For the past 25 years, longitudinal surveys have been an important and valued source of data collection. Of the major NSF-funded surveys, the PSID is primarily longitudinal in nature, and the ANES and GSS both also have components that are longitudinal. Panel surveys have particular value for collecting time-varying information within the same population of respondents, so these data can provide evidence of change in measures over time that are less prone to recall errors and biases. Another benefit of a panel study is that, over several waves, the cost of data collection may be lower than an equal number of similarly sized and equally lengthy cross-sectional surveys.

Researchers are also increasingly taking steps to make cross-sectional surveys more longitudinal. For example, the Integrated Public Use Microdata Series (IPUMS) project at the University of Minnesota has linked records of the Decennial Census to form a longitudinal record, and the American Community Survey has been used as a screener for other studies, suggesting that it might also become an originating source of longitudinal data collection. Furthermore, researchers have periodically linked the groups in the rotating panel design of the Current Population Survey to form a panel structure. These efforts to incorporate elements of panel designs into cross-sectional studies provide an indication of the value of longitudinal data to researchers.

However, longitudinal survey data collection efforts are commonly plagued by respondent attrition from the panel. For example, the longstanding PSID and the newer but similarly structured British Household Panel Survey (BHPS) both display a similar steep decline in response between initial recruitment and the subsequent few waves. In both surveys, the panel attrition hazard rate was above 10% in the first waves before settling between 1-2% attrition per wave for the PSID and 3-4% attrition per wave for the BHPS. Similar results have been observed across many longitudinal panels over time, and the patterns of attrition seem to be consistent with population heterogeneity in innate respondent cooperativeness.

The Educational Longitudinal Survey (ELS), which is a survey of students enrolled in primary school, provides an important perspective on how different approaches to conceptualizing attrition can influence data completeness. The initial wave of the ELS achieved a completion rate of 88% of sampled students. Then, in wave 2, the ELS completed interviews with 91% of the students that responded in the first wave, representing an attrition rate of 9%. Wave 3 completed surveys with 87% of wave 1 respondents, which was only a 4% increase in total attrition. However, the decline in the attrition rate in wave 3 was partially due to an effort by the study to re-contact wave 2 non-respondents and recover data that was initially missed during wave 2.

The approach that the ELS applied of attempting to recover data from wave 2 respondents during the wave 3 interviews enabled them to recover 96% of the wave 2 data. This reveals an important consideration when evaluating how to deal with attrition in longitudinal surveys. Rather than being concerned about respondent attrition from the panel in each wave, the investigators may be better served by focusing on reducing overall data attrition by re-contacting non-respondents from prior waves in an attempt to fill in data that should have been collected in those prior waves.

The National Longitudinal Surveys (NLS), which is a study that collects data from six cohorts of respondents recruited between 1968 and 1997, has similarly experienced significant attrition over time. However, the study has reduced the impact of this attrition by continually attempting to re-contact non-respondents to prior waves and implementing event history calendar methods to fill in the missing data. These efforts have reduced the incidence of missing data substantially, and the success of this approach used by the NLS suggests that a focus simply on respondent attrition may be misguided. Rather, the focus of researchers should be on reducing data attrition by any reasonable means available.

Targeted monetary incentives are one effective approach implemented by the NLS to reduce panel attrition. The differential incentives successfully reduced respondent attrition from the panel, particularly when the largest incentives were targeted at the respondents estimated to have the greatest risk of attrition. This finding supports prior research on targeted incentives as effective tools for increasing response rates but applies it in the context of a panel design.

Interestingly, in the course of experimentally demonstrating the effectiveness of targeted incentives, the NLS also found that overall field costs fell due to fewer interviewer resources being expended to gain respondent cooperation. The decrease in field costs was nearly the same as the additional cost incurred by applying the incentives but nonresponse was decreased among the panel members at the highest risk for attrition, which had significant value since the respondents had been members of the panel for 18 waves. These results were replicated in the NLS panel multiple times, confirming the opportunity for significantly reduced panel attrition over time at minimal additional cost when using targeted incentives.

Taken together, the evidence from the ELS and NLS suggest that some of the concerns about panel attrition may be misplaced, and furthermore that it may be possible to mitigate some the effects of attrition at minimal cost. More fundamentally, these findings indicate that researchers should be more concerned about data attrition than panel attrition, meaning that continued efforts

to re-contact non-respondents from prior waves should be standard practice in order to maximize the potential to fill in the data from missed waves.

Areas for future research:

- Identifying ways to increase the value of panel surveys to researchers accustomed to using cross-sectional data.
- Maximizing the value of having data on people who attrite from panels, as there may be valuable insights for non-response to cross-sectional surveys.
- Evaluation of the problem of the lack of accretion in panels to account for shifting population demographics due to immigration.

References for further reading:

Nekarda, C.J. (2009). A Longitudinal Analysis of the Current Population Survey: Assessing the Cyclical Bias of Geographic Mobility, working paper available at http://chrisnekarda.com/2009/05/a-longitudinal-analysis-of-the-current-population-survey/.

Noah Uhreig, S.C. (2008). The Nature and Causes of Attrition in the British Household Panel Survey, Institute for Social and Economic Research Working Paper, University of Essex 2008.

Olsen, R. (2005). The Problem of Survey Attrition: Survey Methodology is Key, *Monthly Labor Review 128*, 63-70.

Schoeni, R.F., Stafford, F., McGonagle, K.A., and Andreski, P. (2013). Response Rates in National Panel Surveys, *The Annals of the American Academy of Political and Social Science*, *645*, 60-87.

Computation of Survey Weights – Matthew DeBell

When generalizing to a population, weighting survey data can be one of the most important steps prior to analyzing the results. The decisions surrounding how weights are calculated and applied can have enormous statistical and substantive implications for the results that are derived from the data. Weighting is often one of the final steps in survey data generation, and because of this, it is all too often an afterthought for survey researchers. The development and application of survey weights has been the subject of a large body of research within survey methodology, and a number of best practices have been developed over the years. However, there are still a number of key areas where more research is needed to fully understand best practices for weighting and disseminating these best practices to the broad community of survey data users.

Survey weights, at their most basic level, determine how many people each respondent represents. Weights are necessary because, even when sampling is done correctly, there are often unequal probabilities of selection among respondents, and some demographic groups end up underrepresented and others overrepresented in the data. Survey weights balance these

inequalities in the data and enable researchers to transform the observed data such that it is more representative of the target population. In a nutshell, survey weights are applied for the following reasons:

- Determining how many people each respondent represents
- Fixing random error (sampling error)
- Adjusting for unequal selection probabilities
- Adjusting for nonresponse
- Fixing non-random error (within limits)

When applied correctly, weights can reduce bias and allow relatively small samples of data to be used to generate inferences about population values. There are a number of approaches to weighting that have been developed and applied widely, but in general form, the first steps typically include weighting in steps based on selection probabilities first for household and then for person, then nonresponse in observable categories. After this, researchers may post-stratify on key factors (generally demographic) or employ raking or propensity scores.

Weighting is not without its limitations, despite having become widely accepted as a statistically and substantively sound approach to correcting data. For example, weights are not useful for fixing non-coverage error because weights cannot be applied to values of zero. Neither can weights fix extreme nonresponse bias or nonresponse bias for factors that are uncorrelated with the weighting factors. Furthermore, weights cannot be used to correct errors on factors with unknown population benchmarks (e.g. party identification). Lastly, weights do not come without a cost; while they are able to reduce bias, this comes at the cost of increased variance, which thereby increases standard errors and reduces the precision of estimates.

Development and dissemination of widely applicable best practices for weighting is an important area that the existing work on weighting has not covered sufficiently. While the literature on survey statistics covers the statistical theory of weighting in excellent detail, there is very little available to data users about standardizing practice or detailed guides for how to apply best practices when weighting data. This means that survey statisticians are often needed to apply weights correctly and to evaluate their effects on the data and the underlying structure of the data. One area for future research may be to develop a set of rules and procedures that the average data user can implement when applying weights, including a set of standards for when it may be necessary to involve the expertise of a survey statistician.

The result is that most data users are not using survey weights consistently or appropriately; this is largely due to the average survey researcher being unaware of how to apply weights in an appropriate fashion. Even survey statisticians often apply weights in an *ad hoc* manner. This means that the results from analyses are not always transparent, replicable, comparable, or optimal. The clear implication is that more work is needed to identify and disseminate best practices in ways that are accessible to a broader audience than survey statisticians. Thus, there are four key areas for future work:

- More accessible guidance
- Increased transparency of methods

- Clearly replicable methods
- Standardized/comparable methods rather than *ad hoc* approaches

It is impossible to distill weighting down to an extremely simple set of rules and procedures that will be appropriate for all surveys, thus the points above should be viewed as starting points. Identifying "standard" practices does not mean closing off alternatives – rather setting a starting point or frame of reference. There is always more than one way to compute legitimate and effective weights, and these alternative methods each have value, and some may be more contextually appropriate than others. However, this need for flexibility in approaches should not stand in the way of building a set of best practices that can guide the average user toward more appropriate treatment of survey weights.

The American National Elections Study (ANES) provides a good example of how researchers can start to approach this task of identifying and disseminating best practices for weighting data. Leading up to the 2008 ANES, the ANES investigators assembled an expert committee of statisticians and survey methodologists to generate a set of guidelines for best practices for weighting with regard to the particular design features of the ANES. From this set of guidelines, the ANES investigators and staff developed and published a set of procedures codifying these best practices. The goal of these procedures was to describe a general approach to weighting that all users of ANES data or similar data could use; this helped to take the guesswork out of weighting for the average user. The particular procedure recommended by the ANES was a raking algorithm with specific benchmarks and raking factors that were explicitly spelled out in detail. This approach allows the procedure to be standardized for the average user without limiting the flexibility in weighting that might be more appropriate for a specialized user.

But the ANES investigators didn't stop there; they went a step farther and assisted in the development of a weighting tool that could easily be used by any researcher to apply weights and the raking algorithm recommended in the ANES procedure. This tool, "ANESRAKE" is a free package for implementation in the free and open-source 'R' statistical program. It is a practical and automated approach to weighting that enables anyone to apply a generic set of best practices for weighting without needing a course in survey statistics. These are concrete steps that large surveys like the ANES can and should take to make appropriate treatment of survey weights easier for the average user.

However, there are a number of areas where more work is needed. Funding agencies like the NSF can play a key role in moving researchers in a positive direction. For example, by requiring a plan for weighting and a plan for the eventual disclosure of weighting methods to be submitted with grant applications, funding agencies can aid in creating a set of normative standards around the development of transparent and replicable weighting methods. One key point is that future research is needed to identify a set of scientific principles that will guide researchers as they develop and apply weights, thereby moving researchers away from the current practice of researchers using disparate and *ad hoc* methods that are not always fully reported.

Areas for future research:

- Identifying and disseminating a general set of best practices and resources for non-statisticians to use when weighting data
- Improving transparency and replicability when weights are used in practice
- Moving large surveys toward following the ANES model for making weighting accessible to average users

Section 2: Opportunities to Expand Data Collection

Paradata – Frauke Kreuter

Paradata is a term used to describe all types of data about the process and context of survey data collection. The term "paradata" was first used in the survey research context by Mick Couper (1998) to describe automatically generated process data, such as data from computerized interviewing software. Examples of paradata include:

- Listing information:
  - Day, time, edits
- Keystrokes:
  - Response times, back-ups, edits
- Vocal characteristics:
  - Pitch, disfluencies, pauses, rate of speech
- Contact data:
  - Day, time, outcomes
- Interviewer observations:
  - Sample unit characteristics
- Key strokes:
  - Edit failures

Conceptually, these are important data because they allow insights into many sources of total survey error. Some of the most common uses of paradata to examine survey errors include using keystrokes to evaluate measurement error and data validity, or using contact data and observations to examine nonresponse error and adjustment error. These data are distinct from metadata, which are a class of data about the characteristics of the data and data capture systems or "data about the data". Examples of metadata include technical reports, survey instruments, interviewer instructions, show cards, and other documentation of the survey process or variables.

Given the relatively recent discovery of the uses and benefits of paradata, there are not a widely accepted set of best practices for how and when to use all of the different types of information collected as paradata. However, the existing literature does provide considerable information about some specific types of paradata. One of the most commonly used and studied forms of paradata is response time to a question. Current uses of response times tend to be post-hoc and focused on error. For example, examining the characteristics of the survey instrument and setting (Bassili, 1996; Draisma and Dijkstra, 2004; Tourangeau et al., 2004; Yan and Tourangeau,

2008). Factors that tend to increase response times are: poor wording, poor layout, question length, and question complexity. Factors that tend to decrease response times are: logical ordering of questions, respondent practice at survey completion, respondent provision of correct answers, and decreasing motivation on the part of the respondent. Response times have also been used on the interviewer side to evaluate interviewer administration and associated errors (Olson and Peytchev, 2007; Couper and Kreuter, 2012; Schafer, 2013). In the extreme case, response times have been used to identify interviewer falsification when the interview was completed in less time that it would have taken to read each question, much less hear and record the response (Penne et al., 2002). Some novel applications of response times have examined the potential for concurrent use to intervene in self-administered web surveys if respondents answer too fast or slow (Conrad et al., 2007; 2009).

Call record data are another form of paradata that have received considerable attention and research. These data have been used to focus on improving efficiency through identifying optimal schedules for interviewers to reach respondents (Weeks et al., 1980; Greenberg and Stokes, 1990; Laflamme, 2008; Durrant et al., 2010). These data have also been used for identifying potentially important predictors of response. For example, examining the number of contact attempts it took to reach a certain person, when was that person contacted last time, and what is the probability for that person to be at home, or not at home, the next day or the next time you try to reach that particular respondent (Campanelli et al., 1997; Groves and Couper, 1998). Call records have also been used to examine error features such as nonresponse bias analyses and nonresponse bias adjustment (Politz and Simmons, 1949; Kalton, 1983; Beaumont, 2005).

However, despite having examined certain aspects and uses of paradata relatively extensively, there are still important areas for future research to examine. For example, there is still very little research developing systematic approaches to handling keystroke data. This may be due in part to the particularly messy nature of these data, but rather than stymying research altogether, this should be taken as an opportunity for interdisciplinary research that integrates text analysis and survey research on paradata to examine keystroke data systematically. Similarly, there has not been sufficient research on examining face-to-face contact protocol data generated by interviewers after each contact attempt with a household. This lack of research is likely due to the amount of missing data and the very complex hierarchical structure of the data, but this is again simply another opportunity for additional interdisciplinary research to model these data, identify problematic areas and ways to make them more consistent in the future. These types of findings could then easily link back to ways to improve field practice and the practices of the interviewers responsible for neglecting to fill out contact forms or filling them out incorrectly.

Looking beyond the current problems with what is known about uses for paradata, it will be important for future research to examine a couple of important general themes. First, research is needed on ways to enhance real-time use of paradata and the development of best practices around concurrent use of paradata during data collection. Second, research is needed on identifying additional forms of paradata that are available across different modes; there may be additional paradata that could be collected but the opportunities simply haven't been identified or exploited yet. Third, paradata-driven indicators of survey quality (i.e. process quality) need to be explored and developed where they show promise. Other issues that warrant further discussion and development of best practices include the potential for requiring paradata in data

management plans submitted to funding agencies and potential confidentiality issues in the release of paradata.

Areas for future research:

- Expanding the use of keystroke data
  - Development of open-access code repositories
- Development of appropriate statistical methods for handling face-to-face contact protocol data
- Development of better applications for interviewers to use for contact record data
- Identifying approaches to enhance real-time use of paradata during surveys
- Identifying new forms of paradata in different modes
- Development of paradata-driven indicators of survey quality
- Identifying potential confidentiality issues around the release and use of paradata

References:

Bassili, J.N. (1996). The How and Why of Response Latency Measurement in Telephone Surveys. In: Schwarz, N. and Sudman, S. (Eds.), *Answering Questions,* (pp. 319–346). San Francisco, CA: Jossey-Bass.

Beaumont, J.-F. (2005). On the Use of Data Collection Process Information for the Treatment of Unit Nonresponse through Weight Adjustment. *Survey Methodology*, *31*(2), 227-231.

Campanelli, P., Sturgis, P., and Purdon, S. (1997). Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates. Technical report. London, UK: The Survey Methods Centre at SCPR.

Conrad, F.G., Bederson, B.B., Lewis, B., Traugott, M.W., Hanmer, M.J., Herrnson, P.S., Niemi, R.G., and Peytcheva, E. (2009). Electronic Voting Eliminates Hanging Chads but Introduces New Usability Challenges. *International Journal of Human-Computer Studies*, *67*, 111-124.

Conrad, E.G., Schober, M.E., and Coiner, T.E. (2007). Bringing Features of Human Dialogue to Web Surveys. *Applied Cognitive Psychology*, *21*(2), 165-187.

Couper, M. (1998). Measuring Survey Quality in a CASIC Environment. In: *Proceedings of the Survey Research Methods Section,* (pp. 41-49). Alexandria, VA: American Statistical Association.

Couper, M. and Kreuter, F. (2012). Using Paradata to Explore Item-Level Response Times in Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(1), 271-286.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (Para) Linguistic Expression As Indicators of Response Error. In: Presser, S., Rothgeb, J.M., Couper, M.P., Lessler, J.T.,

Martin, E., Martin, J., and Singer, S. (Eds.), *Methods for Testing and Evaluating Survey Questionnaires,* (pp. 131-48). New York, NY: Springer-Verlag.

Durrant, G.B., Groves, R.M., Staetsky, L., and Steele, F. (2010). Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys. *Public Opinion Quarterly*, *74*(1), 1–36.

Greenberg, B. and Stokes, S. (1990). Developing an Optimal Call Scheduling Strategy for a Telephone Survey. *Journal of Official Statistics*, *6*(4), 421-435.

Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York, NY: John Wiley & Sons, Inc.

Kalton, G. (1983). Compensating for Missing Survey Data. Technical report. Ann Arbor, MI: Survey Research Center, University of Michigan.

Laflamme, F. (2008). Data Collection Research using Paradata at Statistics Canada. *Proceedings from the 2008 Statistics Canada International Symposium on Methodological Issues: Data Collection: Challenges, Achievements, and New Directions, Catalog 11-522-X*. Ottawa: Statistics Canada.

Olson, K. and Peytchev, A. (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly, 71*, 273–86.

Penne, M.A., Snodgrass, J., and Barker, P. (2002). Analyzing Audit Trails in the National Survey on Drug Use and Health (NSDUH): Means for Maintaining and Improving Data Quality. International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), Charleston, SC.

Politz, A. and Simmons, W. (1949). An Attempt to Get the "Not-at-Homes" into the Sample Without Callbacks. *Journal of the American Statistical Association*, *44*, 9-31.

Schafer, J.L. (2013). Effects of Interviewer Refresher Training and Performance Monitoring in The 2011 National Crime Victimization Survey. *Center for Statistical Research and Methodology Research Report Series (Statistics #2013-07)*, Washington, DC: U.S. Census Bureau.

Tourangeau, R., Couper, M.P., and Conrad, F.G. (2004). Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly*, *68*, 368–393.

Weeks, M.F., Jones, B.L., Folsom, R.E., and Benrud Jr., C.H. (1980). Optimal Times to Contact Sample Households. *Public Opinion Quarterly*, *44*(1), 101-114.

Yan, T. and Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. *Applied Cognitive Psychology*, *22*, 51–68.

Interviewer Observations – Brady T. West

Interviewer observations are an exciting form of paradata that have recently received more focus among survey researchers. There are two general categories of interviewer observations. The first are observations recorded by survey interviewers for all sampled units that describe selected features of the sampled units, including attempts at recruitment, neighborhood descriptions, and similar observations. The second type of observations are those recorded by survey interviewers for respondents that describe aspects of the survey interview. Observations like: *Did the respondent understand what the survey questions were trying to get at? Did the respondent seem to take enough time? Did they run into cognitive challenges with the actual survey?* In both types of interviewer observations, these can be thought of as observational paradata, or data that describe the process of collecting survey data, that are not automatically generated but rather observed and recorded by the interviewer.

Several large and important surveys already make use of interviewer observations. For example, the Los Angeles Family and Neighborhood Survey (L.A. FANS) study collects interviewer observations on the sampled area for variables such as evidence of crime or social disorder. The National Survey of Family Growth (NSFG) collects data on the sampled household, particularly for interviewer observations of the presence of young children that may not get reported on rosters. The National Health Interview Survey (NHIS) has started to collect housing unit observations related to the health conditions of inhabitants, such as the presence of wheelchair ramps or cigarette butts. The NSFG and Panel Study of Income Dynamics (PSID) both use interviewer observations on the survey respondent for variables such as interviewer opinions of the quality of data provided by the respondent.

There are a number of reasons why interviewer observations are important to collect and study. First, they are an inexpensive and potentially useful source of auxiliary information on all sample units, and this has been demonstrated in face-to-face surveys. Future research should also examine the utility of collecting these data for telephone surveys as well (e.g., interviewer judgments of respondent gender). The second reason that they are important is that prior research has demonstrated that interviewer observations can be correlated with both response propensity and key survey variables, making them useful for nonresponse adjustments and for responsive survey designs (e.g. case prioritization).

However, there may be some drawbacks to attempting to collect and use interviewer observations. For example, not all interviewers view them as easy to collect or worthy of their time and effort. Further, the existing literature on the subject has shown that these observations can be error-prone and are frequently missing. There is also a concern about asking interviewers to essentially become weak proxy respondents for nonresponding households or individuals so that researchers can hopefully learn something about nonrespondents. If the interviewers aren't incentivized to take this task seriously because they would rather focus on completing interviews, then they might not take the time necessary to record high quality information. Lastly, some types of observations (e.g., post-survey) take large amounts of time for interviewers to record, begging the question: do the benefits of the observations outweigh the costs?

Research on interviewer observations has allowed the development of a few best practices. Examples include the recommendation that every observation that interviewers are asked to record should have a specific purpose, such as:

- Nonresponse adjustment
- Prediction of response propensity
- Profiling of active cases for possible targeting
- Assessment of data quality

Collecting observations with no *a priori* purpose is likely to be a waste of interviewer time and researcher money. Furthermore, observations collected on all sample units should be correlated with key variables and/or response propensity. This suggests that one opportunity for future research is to focus on identifying and understanding these associations. Once these associations are empirically established, then the observation requests to interviewers should ideally be designed as proxies for key measures in the survey so that they can act as a form of validation or data quality check. This would drastically increase the benefit that researchers get from interviewer observations.

Once interviewer observations have been collected, it is important that the survey organization *actually analyzes* the observation data. Even if the data are of questionable quality, there may be important information that could point toward operational improvements that the organization can make. More specifically, these observations may point to problems with the questionnaire in general or indicate data quality issues that should be addressed. As part of this point that observations should be analyzed, when possible, the quality of observations should be assessed using validation data. The easiest validation might be the actual survey reports of the same phenomena observed, but there may be opportunities to validate using administrative records or even observations by another interviewer.

Methods for standardizing the ways in which observations are collected is another key area where best practices have been developed and implemented but further research may be warranted. For example, concrete training approaches such as using visual examples of how to make effective observations should be implemented as part of interviewer training. An example of this can be found in the European Social Survey (ESS), and the NHIS has started to employ this kind of training. Interviewers should also be provided with known and observable predictors of features that they are being asked to observe; this approach has been implemented by the NSFG. Lastly, interviewers may be asked to provide open-ended justifications for why they record a particular value for an observation; this is also an approach used by the NSFG.

However, there are several areas where future research is needed before additional best practices can be defined. First, more work is needed to assess the validity and accuracy of interviewer observations across a variety of different face-to-face surveys. Second, future research is also needed to identify correlates and drivers of observation accuracy, such as features of interviewers, respondents, or areas that can predict interviewer observation accuracy or inaccuracy, or alternative strategies used by interviewers that can lead to higher accuracy.

Third, future research needs to examine what interviewer observations add to existing auxiliary variables such as those in commercial databases. It is important to know if the observations explain additional variance in survey outcomes on top of existing variables. Fourth, the statistical and operational impacts of observation quality on responsive survey design approaches should be examined to determine if decisions based on the observations (e.g. targeting certain cases) backfire at a lower threshold of quality or if the observations serve to improve efficiency regardless of quality.

Fifth, additional research is needed to evaluate the ways that error-prone interviewer observations affect statistical adjustments for survey nonresponse; only weighting adjustments have been studied thus far, meaning that there is a need for further research. Sixth, research is needed on effective design strategies for improving the quality of interviewer observations; for example, does providing interviewers with known auxiliary predictors of features improve the accuracy of their observations or not?

Seventh, work is needed to understand how post-survey observations might be used to improve survey estimates. For example observations could be used to design calibration estimators or as indicators of data quality to inform researchers about which cases to consider dropping due to poor response quality. Eighth, more work is needed to identify and understand the sources of interviewer variance in observations and observation quality. For example, qualitative studies of different strategies used to record observations in the field and interviews with interviewers to determine what their actual practices are. Lastly, and perhaps most importantly, it will be important for future research to identify and measure the empirical trade-offs between the costs of collecting interviewer observations versus improvements in survey estimates from collecting the observations.

Areas for future research:

- Assessing the validity and accuracy of interviewer observations
- Identifying correlates and drivers of observation accuracy
- Determining the added value of observations over other existing auxiliary variables from commercial databases
- Examining the statistical and operational impacts of varying observation quality
- Understanding the effects of potentially error-prone observations on statistical adjustments for nonresponse
- Designing optimal strategies for improving the quality of interviewer observations
- Identifying ways to use post-survey observations to calibrate data or give indications of data quality
- Understanding sources of variation in observations and observation quality between interviewers
- Identifying and measuring trade-offs between costs and benefits of collecting interviewer observations

Leave-behind Measurement Supplements – Michael W. Link

Leave-behind measurement supplements, as the name implies, are surveys or survey-related tools left with respondents after an interview has been completed. These tools are frequently used by research organizations but have not been the subject of much empirical research. Much is known about the components of leave-behind materials but very little about the methodology and best practices, and not much has been published on this subject despite the widespread use of the approach.

Leave-behind measures have a number of defining characteristics. First, they almost always involve self-administration. Second, the data collection mode is often different than the initial mode, meaning that when they are used, leave-behinds are often components of a multi-mode design. The type of data that these measures provide is typically supplemental for a study, but on rare occasions they have been used to collect the primary data. Another unique feature is that the leave-behind task is completed after the end of an initial survey or interview, meaning that this is a multi-stage (not simply multi-mode) approach to data collection. Leave-behinds are nearly exclusively implemented by large surveys that often involve complex data collection efforts; rarely are they part of data collection with smaller or more straightforward studies. Leave-behinds may take many forms, though additional surveys, diaries, electronic monitors, and physical specimen collection devices are most common. Diaries are perhaps the most popular of the leave-behind methods in survey research, and The Nielsen Company's "people meters", which are used to generate television ratings, may be the most prominent example of the method being applied and used as the primary source of data collection. Not included in this definition of leave-behind measures are traditional mail or panel surveys, self-initiated online surveys, or Audio Computer-Assisted Self-Interview (ACASI) segments conducted during a larger face-to-face interview.

There are many different data collection purposes that leave-behinds are suited for. At the most basic level, they are useful for collecting more information to expand the initial data collection effort, but they are also useful for reducing respondent burden by allowing them to complete a portion of the data collection on their own schedule. They are also very well suited for sensitive topics or when privacy concerns may be an issue; in these contexts they may lead to less social desirability bias and higher quality data. For some types of data collection, using a leave-behind may provide an opportunity to improve data quality by reducing the need for respondents to recall things such as daily activities that are more accurately captured by use of an activity diary left with the respondent after the initial interview.

While these methods have been widely applied in practice, there is an apparent lack of empirical research on the methodology from which best practices might be derived. When leave-behinds do appear in the literature, they tend to be discussed as a component of a broader study and not the specific focus. This may be because leave-behinds are perceived to be adjunct data and not the primary focus of the study; thus they are not explicitly examined or the findings regarding the leave-behinds themselves are not reported widely. This presents an important opportunity for future research to examine fundamental questions about how these methods are being used, how effective they are, and how they can be improved.

One important and exciting area for future research is identifying ways that new technologies can enable new and improved methods of leave-behind measurements. Mobile platforms, apps for tablet devices or smartphones, and other new technologies offer innovative approaches to this suite of methodologies that may expand their applicability and utility for surveys. One of the appeals of incorporating these new technologies, as part of the leave-behind approach, is that they could be used to facilitate quick and easy communication with respondents rather than expensive and potentially intrusive face-to-face or telephone contact methods. These technologies could enable respondents to be prompted to enter data or to upload their data for researchers to analyze. Furthermore, GPS enabled devices have become ubiquitous, and if respondents consent to being tracked using their smartphone or another device, this could allow incredibly rich validation data to be collected in conjunction with time-use diaries and other self-report methods. Bluetooth-enabled devices could similarly revolutionize data collection as they enable researchers to passively and semi-passively capture a wide array of respondent activities; for example, it is possible to passively record blood glucose, blood oxygen, and pulse data passively using Bluetooth devices. Other novel opportunities for collecting data using new technologies include image or video collection, audio capture, and text-entry.

Beyond exploring the promise of new technologies, additional research is needed at a more fundamental methodological level. For example, what lessons can be drawn from focusing on leave-behinds as a distinct methodological approach? Can generalizations be made across different approaches, or are the techniques used in practice too varied to allow for comparison? How do data quality concerns around leave-behinds differ from "primary" modes of data collection? Do leave-behinds actually reduce respondent burden? What respondent compliance concerns are associated with leave-behinds? Is satisficing behavior influenced by leave-behinds? Specifically regarding data quality issues, do timing and context change responses that might have been obtained otherwise? Are data collected with leave-behinds comparable with other forms of data?

Areas for future research:

- Applications of new technologies to improve and expand leave-behind measurement
- Examinations of leave-behind measurement tools as distinct methodological approach
    - Generalizing across different techniques
    - Data quality concerns unique to the methodology
    - Impacts on respondent burden
    - Issues surrounding respondent compliance
    - Effects on data quality in terms of measurement error when compared with other methodologies

Experience Sampling and Ecological Momentary Assessment – Arthur A. Stone

Ecological momentary assessment (EMA; also known as Experience Sampling) is a family of measurement techniques where respondents, who have previously been recruited to be part of a panel, are contacted according to a predetermined schedule and asked to report information about their current state. This may include questions about the psychological state of the respondent,

their experiences, behaviors, symptoms, features of their environment, and even physiological metrics. The goal of these methods is to capture data on experiences and behavior as precisely as possible. The gains in accuracy stem from the ability to capture daily experiences and behavior with minimal recall bias or respondent forgetting; this enables researchers to generate high levels of ecological validity or correspondence with what respondents actually experienced.

EMA methods allow researchers to study time usage in more depth and with greater precision than many other approaches allow. Using these approaches, researchers are able to study patterns of experience or behavior, particularly those that fluctuate rapidly like emotions or symptoms and are harder to recall later. EMA enables research on within-day contemporaneous and lagged associations between experiences and behaviors; for example, how does having a cigarette craving associate with the behavior of having a cigarette? These methods have also been used to link experiential and behavioral data with real-time physiological data such as blood glucose levels, electrocardiogram (EKG), electroencephalogram (EEG), and other cardiovascular measures.

Conceptually, EMA methods are an important class of measurements because they allow access to information that can be challenging to capture using other approaches because of how and where they are stored in memory. Immediate information is stored in experiential memory, which captures information about what respondents are experiencing at the moment. This type of memory tends to be very short-lived, making it hard to access later such as when a more traditional survey might ask about the experience. Episodic memory captures memories of experiences, whereas semantic memory stores a different kind of information, typically things like attitudes and beliefs about experiences. When a recall period expands from *"How are you doing right now?"* to "*How are you doing over a day or over a week or over a month?"*, there is a shift in the kind of memory that respondents access in order to create the answer. As the memory period increases because the recall period increases, respondents' answers tend to shift to semantic memory and beliefs. So in the case of EMA, rather than attempting to measure beliefs about things that happened, the goal is to measure the experience itself.

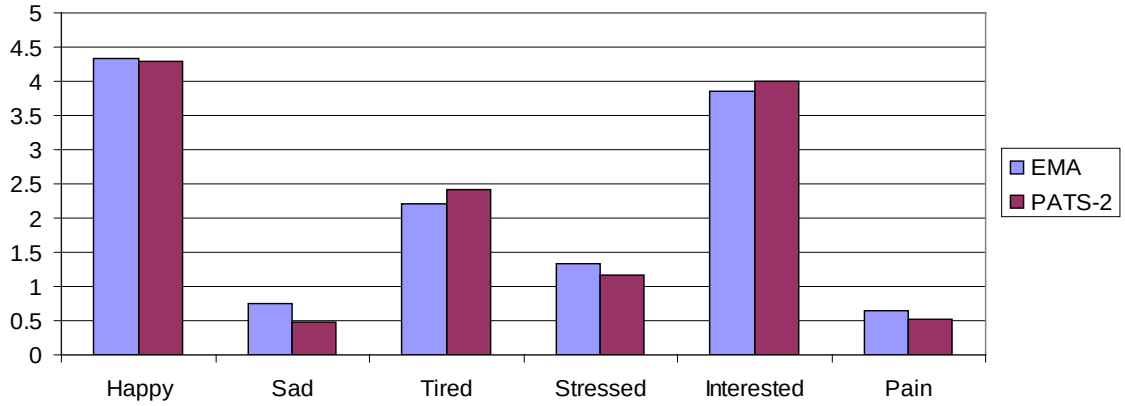
Researchers who apply EMA methods are typically concerned about distortion associated with long recall periods, which may include periods as short as a week or even a day. There are a variety of cognitive heuristics that the brain uses in order to summarize information in ways that may reduce response accuracy over longer recall periods. A key point is that respondents are not aware heuristics when they activated. For example, a well-known heuristic is the "peak-end" rule, which describes people's tendency to remember peaks of experience (things that are salient) and things that are relatively proximal to when they are completing the questionnaire. A similar heuristic appears when respondents report current levels either as a proxy for the recall period asked about or to alter their response by reporting the past experience relative to the current one. Essentially, these issues lead to the conclusion that asking people about certain kinds of experiences, symptoms, or behaviors over relatively long periods may be fraught with bias. These have been recognized for a long time, including by survey researchers (Bradburn et al., 1987). However, EMA methods are just beginning to be used in the broader survey research community.

While ecological validity isn't typically a major concern for survey researchers, there are some particular contexts in which it is an important consideration. For example, in health studies or drug trials, it is often desirable to capture data on what actually happened to respondents (symptoms, side-effects, etc.) rather than the belief-biased responses that respondents may provide later. The underlying concept is that to understand the experiences that people are actually having in the most valid way, researchers need to representatively sample situations from the relevant universe of situations that includes the experiences of interest. The approach then is to contact respondents at a representative set of random intervals and ask them about periods of time immediately preceding the contact.

The features of EMA have a number of benefits and drawbacks that are important for researcher to consider before applying the methods. In terms of real-time data capture, the primary benefits are that recall bias is significantly reduced or eliminated because the sampling period occurs concurrently with or immediately following the experience. It can also allow a window into daily patterns and rhythms of experience for respondents. However, the drawbacks of real-time data capture are that it captures only point estimates rather than more global evaluations, the sampling framework is complex and challenging to implement, important events that occur may often happen outside of the very short window of the recall period, and lastly, the approaches tend to be expensive and burdensome for respondents.

There are a number of concepts that apply to typical questionnaire research that EMA researchers have identified as best practices for reducing bias due to recall errors on the part of respondents. First, it is typically best practice to limit the recall period to a very short amount of time immediately preceding the contact with the respondent. A second technique is to elicit reconstruction of the recall period by respondents. Third, researchers can use very precise questions to ease the recall process for respondents, and similarly, researchers can limit queries to information that can easily be recalled, such as salient events. For example, EMA methods are useful for collecting in-depth information about a single day or a portion of a day but have limited value for collecting data over longer recall periods. Furthermore, the burden on respondents participating in EMA studies is sufficiently high that long-term panel participation with high levels of daily measurement is infeasible. For some surveys this approach may be valuable, but others might find it untenable.

The Day Reconstruction Method (DRM) was developed in response to some of these challenges with EMA approaches. The DRM was designed to allow researchers to reproduce the results that would have been achieved by EMA through having respondents systematically reconstruct the recall period (Kahneman et al., 2004). The American Time Use Survey (ATUS) and the Princeton Affect and Time use Survey (PATS) have both implemented this approach successfully. Preliminary research has indicated that the DRM produces substantively similar results to EMA, as seen below:



However, the DRM method requires a technologically sophisticated administration approach, and it is still time-consuming for respondents, often taking 30-45 minutes. Thus, while the DRM approach is an improvement in terms of its applicability to surveys, future research is needed to create versions of the DRM task that are more amenable to the survey context. For example, versions of the DRM that are implemented over the Internet are now available in addition to paper-and-pencil and interviewer-administered versions.

In summary, there is considerable interest in characterizing daily experiences and behaviors in real-time or near real-time. Many subject areas studied with survey data could potentially benefit from these kinds of approaches, but more research is necessary to identify optimal ways to integrate these approaches with large-scale data collection operations. This is not to imply that surveys should move to begin collecting these data; given the logistical and cost challenges with collecting EMA and DRM data, there needs be a clear rationale for attempting to collect detailed daily information in large-scale surveys. Several studies have been successful with alternatives to EMA, which is promising because it makes these types of data more feasible to collect in conjunction with more traditional survey data. However, it is not yet entirely clear how well DRM approaches replicate EMA, and while there is promising recent evidence, more research is needed in this area.

Areas for future research:

- Identifying large-scale surveys that could benefit from EMA or DRM data and conducting feasibility testing
- Determining best practices for implementing EMA or DRM methods in conjunction with traditional survey data collection
- Continued research comparing EMA and DRM methods
- Identifying ways to reduce the time and costs associated with DRM methods
- Identifying novel ways to leverage new technologies to collect these data at lower cost and in larger quantities
- Measuring respondent burden when these methods are applied
- Addressing data consistency and quality concerns

References:

Bradburn, N.M., Rips, L.J., and Shevell, S.K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science, 236*, 157-61.

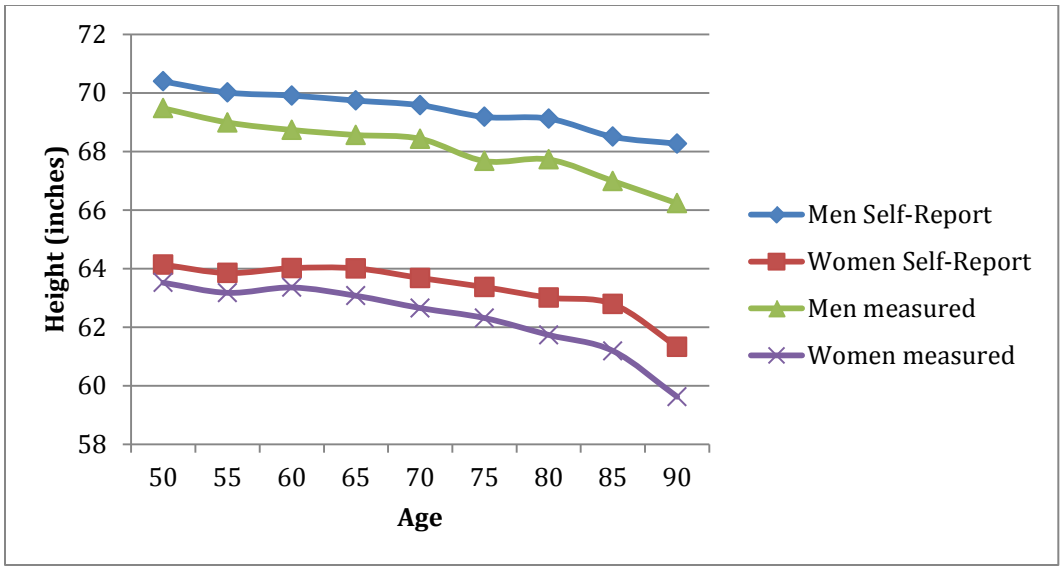
Kahneman, D., Kreuger, A.B., and Schkade, D.A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science, 306*, 1776–1780.

Biomarkers in Representative Population Surveys – David Weir

Biomarkers are a class of measures that are collected via physical specimens provided by respondents. These are typically direct measures of biological states, including disease, physiological functioning, and physical traits such as respondent height and weight. Prior to implementation in health surveys, these types of measures were typically collected in the clinical context and often only using convenience samples rather than representative samples of the population. Now there are at least four large nationally representative surveys currently collecting biomarker data: National Health and Nutrition Examination Survey (NHANES), Add Health, National Social Life, Health, and Aging Project, and the Health and Retirement Study (HRS). Of these surveys, NHANES is considered the gold standard, as it is essentially a study designed to collect biomarker data on large and representative samples of the population.

There are several different types of biomarker measures that are commonly collected, as was mentioned above. These are typically minimally invasive and range from physical measurements of height and weight to biochemical assays from blood, other fluids, or body parts. Occasionally, advanced imaging technology is involved such as X-ray or MRI, but this is far less common. Lastly, a relatively new measure that is expected to have great utility in the future is DNA samples.

Biomarkers are valuable measures to collect for a number of reasons. First, they provide objective measures of health that are more accurate and less subject to bias than self-report data. They are also able to measure biological traits and states that respondents themselves are often unaware of during the survey. For example, as can be seen in the figure below, the HRS compared self-reported height with measured height and found that, across ages and gender, measured values were consistently lower than self-reported values.



Second, biomarkers enable researchers to generate descriptive statistics about the health of the population. Third, researchers are able to use direct measures of health as dependent variables to identify important predictors of specific health states. For example, in epidemiology there are theories that social economic status affects health outcomes through stress-related pathways; biomarkers of those systems allow researchers to test that mechanism directly. And lastly, economists use objective health metrics as predictors of other important variables such as productivity and economic outcomes.

There are a number of cautionary points and best practices for biomarker collection that have been indicated in prior work. The first is that participation in biomarker collection may be related to the health state that the researcher is trying to measure, meaning that nonresponse bias is a concern for biomarker measures. This can be addressed by pairing physical measures with self-report items because it allows imputation or re-weighting solutions to capture the full range of variation in the biomarker. Second, researchers need to be aware of the effects of applying different cut-points to the data, meaning that slight differences in where cut-points are assigned may have significant substantive effects on the results obtained. For example, Body Mass Index (BMI) is calculated by dividing weight by height-squared. Weight estimates tend to be fairly accurate, but the errors in self-reported height above will be magnified in self-reported BMI estimates. The result is that self-report BMI is about 4% low on average with 29.5 percent of respondents being classified as obese. But the mean of BMI is very close to the cutoff for being obese, and measured BMI indicates that 38.2 percent of respondents are obese. This means that the 4 percent error in self-reported BMI translates into a 29 percent increase in the fraction of the population that actually qualify as being obese when BMI is measured using biomarkers. So the cut-off for being classified as obese has significant substantive implications for how both the self-report and biomarker data are interpreted.

The third best practice relates to interviewer training. There is a range of concerns that arise here, from addressing respondent confidentiality concerns to maintaining a sterile environment when biological specimens are being collected and transported. Properly training interviewers to handle the unique demands of collecting biomarkers is important. Preliminary evidence indicates

that interviewer uncertainty and unfamiliarity with collecting biomarkers may be associated with lower levels of compliance among respondents. This implies that biomarker collection cannot simply just be added onto an existing survey without investing in significant training efforts to not only teach interviewers how to collect the biomarker data but also to make them comfortable doing it.

It is important to note that biomarker collection has two potentially important impacts on surveys that must be considered before implementing these methods. First, collecting biomarkers is very time-consuming. This drives up respondent burden and means that interviewers are able to collect fewer interviews in the same amount of time. Second, and relatedly, the amount of effort and cost that must be put into actually recruiting respondents and conducting the interview also increases. The good news is that, when the additional effort is invested in recruitment, response rates seem to not be affected by the addition of biomarker data collection. However, there do seem to be significant racial differences in respondent cooperation specifically and uniquely with regard to biomarker data collection, indicating that further research is needed into how to overcome this potential nonresponse bias.

Two recent developments in biomarker collection promise to vastly expand the breadth and depth of biological information that can be captured by health surveys: dried blood spot and DNA samples. Dried blood spots (DBS) are easily collected by using a small lancet to prick the respondent's finger and then collecting very small drops of blood on filter paper where it can be stored in dry conditions. In addition to the ease of data collection, DBS enables collection of a blood-based biomarker by regular interviewers, that is, interviewers without the phlebotomy training and certification that is required to draw whole blood from respondents. Storage and handling of DBS is also considerably easier than with whole blood, which requires careful temperature control and rapid processing after collection. DBS also has the advantage of being much cheaper to collect and process than whole blood.

The tradeoff that comes with DBS is that there are a limited number of biological assays that can be used to analyze the DBS, meaning that a smaller range of analyses can be performed relative to whole blood. There are also some concerns about the quality of these measures. This is an area where future research is needed; increasing the range of analyses that can be performed on DBS and improving the quality of the measures will help this method to achieve its promise as a key innovation in biomarker research. Lab validation studies comparing within-subject DBS with whole blood samples, test-retest reliability of DBS assays, and comparisons of population distributions of estimates attained between studies using whole blood and DBS are still needed.

The second recent development in biomarker research that holds great promise for the future is in the area of genetic biomarkers. In recent years, collecting DNA from respondents has become very easy and inexpensive, with a number of vendors producing cost-effective and easy to use kits that regular interviewers can use during interviews. Currently analysis costs are still high, and the range of useful analyses that can be conducted on DNA is relatively low; however the costs are dropping rapidly and analyses are becoming increasingly useful. These are minimally invasive tests typically only involving a saliva sample, meaning that field implementation is also relatively easy for interviewers. The DNA analyses approaches still require significant future

research before the potential of genetic biomarkers as part of survey data collection is fully realized, but this is arguably a major component of the future of biomarker research.

Respondent consent, confidentiality, and notification guidelines and concerns are other important areas requiring future attention. Researchers need to define the ethical framework for how to ensure that respondents are aware of the implications of their consent and how the data will and will not be used. For example, most researchers do not notify respondents if biomarker data indicate disease or health risk factors; this is an area where the ethics of handling these unique types of data have not been fully explored and best practices defined. Confidentiality is also a concern, as genetic data and data from other biomarkers can be extremely sensitive and identifiable. Continued discussion of best practices for respondent confidentiality is necessary as biomarker data become increasingly powerful and prevalent.

In summary, applications of biomarkers in survey research have only scratched the surface of their potential. The existing methods have demonstrated the feasibility of combining biological data collection with population surveys, but further research is needed to develop the tools even further to increase the amount of data collected, the quality of those data, and the utility of those data for analysts. New technologies and methods will reduce costs and continue to drive biomarker research into new promising areas of population health research, but more work is needed to fully maximize the potential of these approaches.

Areas for future research:

- Identifying approaches to reduce the nonresponse bias associated with race in response to biomarker requests
- Methods for improving the quality of biomarker measures
- Improving interviewer training to ensure data quality
- Expanding comparisons of similar methods such as whole blood and DBS to identify areas of inaccuracy
- Testing and implementation of new technologies for collecting and analyzing biomarker data
- Maximizing the potential of newly cost-effective biomarker data such as DNA

Specialized Tools for Measuring Past Events – Robert Belli

Respondents to surveys often provide retrospective reports of events or behaviors that have high levels of error, indicating that the responses are often not reliable and raising questions about the quality of the data being collected. Survey researchers have responded by developing a number of methods and tools for maximizing the quality of these retrospective reports. These innovations have led to number of best practices being developed but there is still a need for improvement through continued research on developing new methods and improving those currently being used.

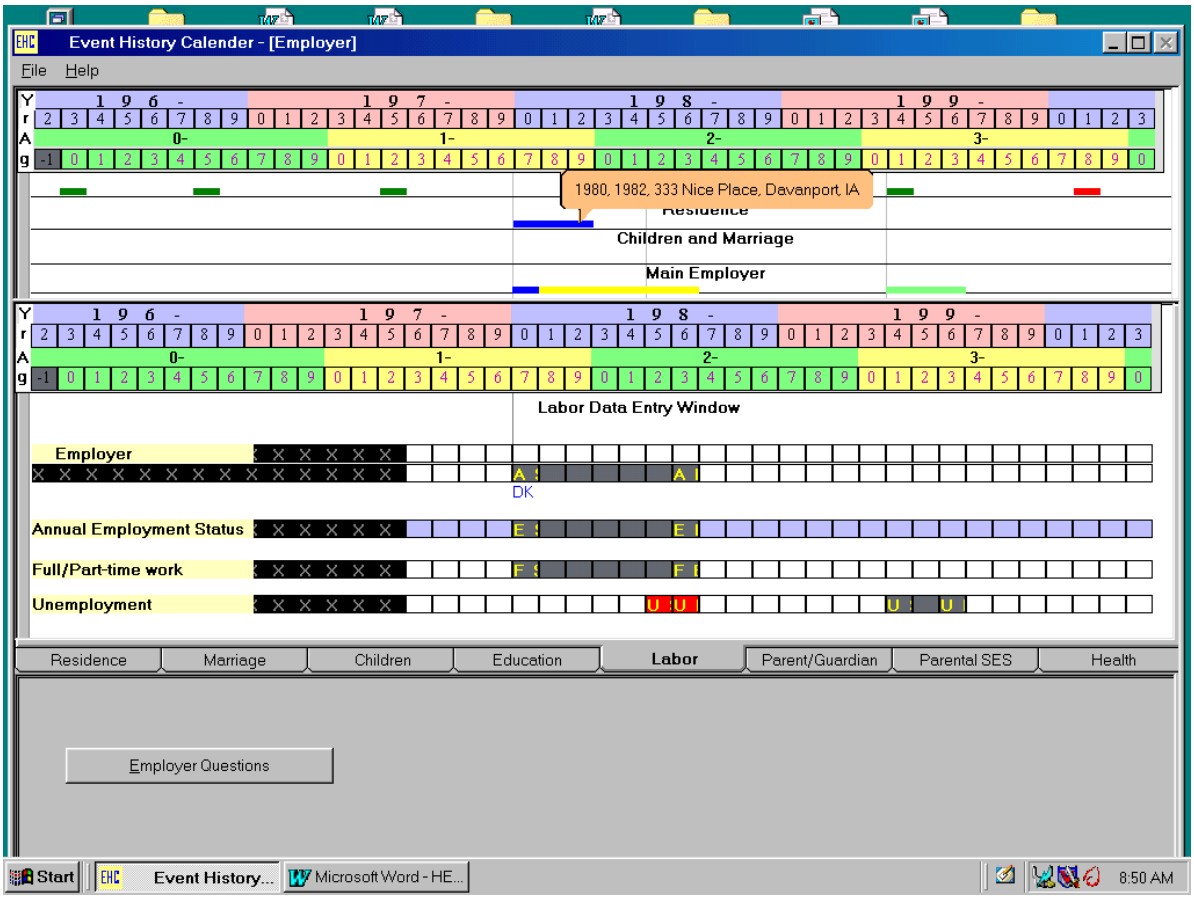
In conventional interviewing practice, one important goal is to measure only the variance in respondent reports, and this often means minimizing other exogenous factors that could influence these reports, such as interviewer effects. One important method toward reducing these

exogenous effects has been the development of standardization practices so that respondents are all exposed to the same stimuli prior to providing their responses. However, standardization does little to help with autobiographical memory recall, so other methods were developed including calendar interviewing and time use diaries.

Calendar interviewing refers to the use of event history calendars that are used as visual aids to assist respondents with recalling with greater precision when a particular event occurred. This method typically includes using a physical calendar as a visual aid for the respondent to improve the precision of their estimates. It is worth noting that calendar interviewing is not generally compatible with standardized interviewing because the steps of the process will look different for each respondent, even if the procedure is the same.

Underlying this method is the idea that one of the best ways to get respondents to provide more accurate reports, in terms of reconstructing their past more accurately, is to use whatever information the respondent can provide from memory as cues for remembering other thematically or temporally related events. For example, a respondent may tell the interviewer that they had a child in June, 1984 and then the interviewer will use that event and date as an anchor to determine when other events in that time period occurred. These are often events that are harder to remember, such as the date that the respondent moved to a new home.

Because most of the cues collected and used for calendar interviewing are idiosyncratic to individuals, standardization is not a practical approach when using the event history calendar method to improve recall accuracy. This means that flexible interviewing approaches that emphasize conversational interviewing are typically used when calendar methods are being used. Calendar methods are often used to assist with collecting data over very long reference periods such as years, decades, or the entire lifespan of the respondent, as can be seen in the example calendar computer-assisted personal interviewing (CAPI) interface below.



Interviewers are trained to apply at least three types of memory retrieval strategies using the cues they collect with the calendar method. First, is the sequential retrieval strategy, in which the interviewer helps the respondent to identify a salient event and then uses that event as an anchor to move forward or backward in time to identify other temporally proximal events. An example of this would be asking a respondent to reconstruct their employment history forward through time starting from the start of their first job. Second is parallel retrieval in which the interviewer uses a cue provided by the respondent to identify other events that co-occurred with the cue, for example, having lived in a particular place could act as a cue for who the respondent's employer was at the time. Lastly, there is top-down retrieval which identifies a general event and then drills down to collect more specific information. For example, after having identified an employer for a specific period, the interviewer can then probe about whether or not there were any changes in income.

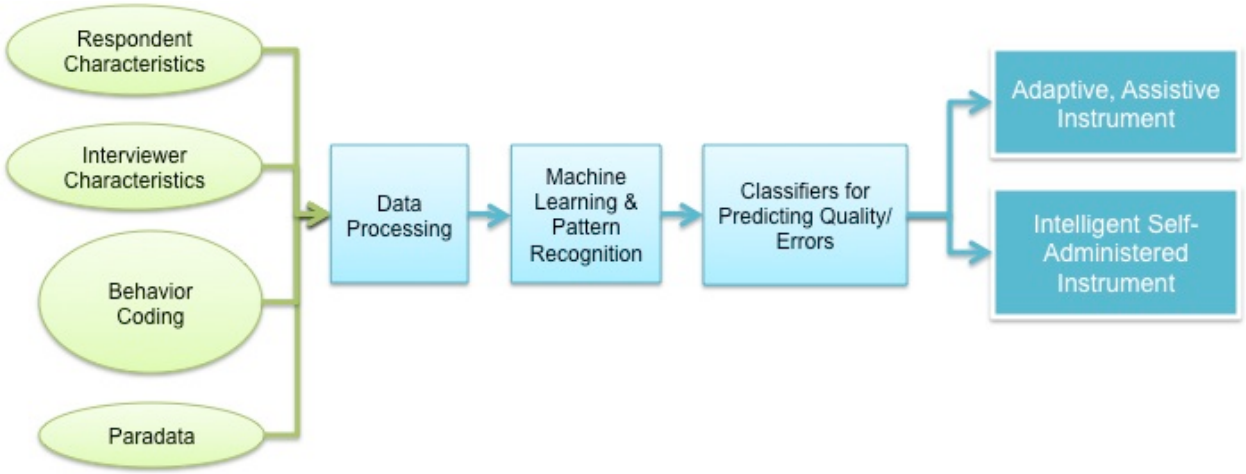
Existing research on the results of the calendar method have indicated that it generates more accurate reports, particularly for temporally remote events, than conventional questionnaires alone. Reports of sensitive events also seem to be more accurate when the event history calendar is used than a conventional questionnaire alone.

Time-use diaries are also commonly used for improving respondent recall accuracy. In many cases, these are self-administered by the respondent; however, they can also be used by

interviewers to aid respondent recall of events from the past day or week. In either context these tools have been shown to increase recall accuracy of past events. These and similar approaches are discussed in more detail in the "*Leave-behind Measurement Supplements*" and "*Experience Sampling and Ecological Momentary Assessment*" sections elsewhere in this report.

However, these measurement tools are costly to implement due to the amount of interviewer time they require and the potential of adding burden on respondents. Future research is needed on ways to computerize more of these tools and make them adaptive and intelligent. Moving toward more self-administration with computerized instruments that are easy to use and capable of collecting data at the same level of accuracy and precision as interviewer-administered tools would drastically increase the use of these tools.

There are many sources of data that future computerized tools can draw on and methods for making use of these data to develop better interfaces and tools for self-administered questionnaires. However, much more research is needed before these ideas can become reality. Below is a general model of easily collected data (in yellow ovals), existing methods for processing these data (in light blue squares), and theorized tools (in dark blue squares) that could be developed to ease implementation of specialized measurement tools for past events.



Once systems like this are developed, a considerable amount of research will need to be done to evaluate the reliability and validity of the data collected when compared to traditional methods. Effects of self-administration on accuracy, comprehensiveness of data collection, and compliance will be of particular concern. The exploration and development of on-the-fly data quality metrics will also be important to making smart survey instruments feasible.

In summary, specialized tools for measuring past events are uniquely capable of improving respondent recall accuracy for events that occurred years or decades in the past. Event history calendars have demonstrated superior performance to conventional questionnaires, indicating that these tools do work. Time-use diaries have similarly been shown to improve recall of episodes from the previous day or week when compared with conventional questionnaires. These methods have drawbacks in terms of cost and the amount of time that they add to the interview process, but future research on computerizing these tasks and developing smart survey

instruments could make it possible to collect highly accurate and precise data using self-administered modes.

Areas for future research:

- Identifying best practices for reducing the time associated with event history calendars and time-use diaries
- Developing and testing computerized systems to explore whether self-administration is feasible for these tasks

Section 3: Linking Survey Data with External Sources

Linking Survey Data to Official Government Records – Joseph W. Sakshaug

Government agencies, medical facilities, and market researchers are among the entities that collect vast amounts of data in the form of administrative records. These records are rich sources of data that can potentially be linked to survey data in valuable ways. Official records such as government documents, medical records, or academic transcripts can not only provide a source of supplemental data but also act as a gold standard to validate the accuracy of self-report data from surveys.

Administrative records linkage refers to appending datasets based on one or more linking variables that exist in both datasets. These linking variables could be at the person, household, or establishment level and could be such variables as names, addresses, Social Security numbers, tax identification numbers, or other variables that can reliably associate disparate datasets.

There are a number of reasons that linking administrative records is important. First, as was mentioned above, are the methodological purposes of assessing data accuracy and the reliability of self-report data from surveys. Another methodological benefit is that these data provide a way to assess nonresponse bias by comparing the survey data collected from respondents with the record data from non-respondents. There are also substantive benefits to linking administrative records with survey data. For example, it can allow for longitudinal analysis because many types of records are in time-series form such as medical or tax records. Linking also permits researchers to investigate complex policy-oriented questions such as trends in healthcare spending among older populations or lifetime earnings and retirement planning.

Government records are some of the most important types of records that are commonly linked to survey data. Popular administrative databases include Social Security records, which contain detailed earnings and benefit histories, and Medicare claims records, which document Medicare enrollment and detailed healthcare expenditure records among Medicare beneficiaries. Last is the National Death Index, which is a database that collects death certificate records from the vital statistics offices of each state; these are aggregated by the National Center for Health Statistics and made publicly available.

Three approaches to linking are commonly used. The first is *exact linkage*, which involves linking administrative records to survey data using a common variable that acts as a unique identifier. These unique identifiers are things like Social Security numbers, Medicare numbers, or tax identification numbers. Respondents are typically relied upon to provide the unique identifier and must also provide informed consent before the linkage can be made. There are some practical concerns associated with attempts to do exact linkage, including the fact that consent rates vary across studies and sets of records. This can lead to reduced sample size and biased inference if the people who do not consent are systematically different from those who do on some unmeasured dimension.

The second approach is *probabilistic linkage*; this can be used when there is no unique identifier on which to match the datasets. In this case, there are other potential identifiers that can be used together to link records with a certain probability of accuracy. Commonly used identifiers are names, dates of birth, and addresses. These identifiers are used to calculate the probability that an administrative record and a survey report belong to the same unit with the match status being determined by a pre-specified probability threshold and decision rule. This approach is commonly used by government agencies such as the Census Bureau and the Centers for Disease Control and Prevention with both of these organizations having developed specialized software packages designed to implement the procedures. Some of the practical issues surrounding using probabilistic linkage are that it is very difficult to estimate the frequency of false matches and false non-matches. Furthermore, the matching variables may have varying levels of accuracy or missing data themselves, which can add more uncertainty. Lastly, linking three or more databases can be very problematic, and there are no accepted best practices for doing so.

The third approach to linking administrative records to survey data is *statistical matching*. This is another method that is used when exact record linkage is not possible. Statistical matching takes the two datasets and uses statistical models to predict the similarity between records and attempts to merge the datasets without regard for identifying cases that actually belong to the same unit. Essentially, this approach uses variables that the two datasets have in common to link statistically similar records; these matching variables may be age, gender, race, and other similar socio-demographic variables. Metrics such as Euclidean distance, predictive mean matching, and propensity score matching can then be used to identify similar records in each dataset and match them. Statistical matching has many practical problems that must be addressed when it is being implemented, but one of the most basic issues is that it makes very strong statistical assumptions (e.g. the conditional independence of variables Y and Z, given common variable X) that are difficult to test with the actual data and therefore may be unjustified. Considerably more research is needed to evaluate whether statistically matched records actually reflect true values in the population.

Despite having developed these important and useful approaches to linking administrative records data with survey data, there are still many opportunities for basic and important future research on this subject. At a very fundamental level, research on how the properties of these different linkage types influence data quality is still needed. It is unclear how linkage errors affect subsequent statistical analyses or whether the strengths of one technique can be used to overcome the weaknesses of another. For example, exact matching and probabilistic linkage require informed consent from respondents, whereas statistical matching does not; under what

conditions does this make statistical matching preferable over the additional effort and potential biases involved with obtaining informed consent? Also on the subject of consent, how low do consent rates need to get before researchers begin considering alternative (non-exact) approaches to linkage? What are the theoretical mechanisms that drive the linkage consent decision? How does this decision differ from the decision to participate in a survey? And how should researchers balance the tradeoffs between data utility and data confidentiality in the unique context of linking administrative records? These are all important questions for which future research is needed.

Methods for linking three or more data sources simultaneously are currently imperfect and in need of additional research. The most common method is called "chaining", where data sources are linked sequentially starting with the most reliable data source. Very little is known about how well this approach works or whether or not it is an optimal method. More evaluation is needed on how linking multiple data sources may affect subsequent statistical analyses.

Lastly, one idea that is starting to gain some attention is the notion of starting with the administrative records data and designing the survey around them. This approach turns the traditional view of linking records on its head and treats the official records as the primary dataset and the survey data as supplemental. This approach could lead to reduced data collection costs, reduced respondent burden, more efficient survey design and use of records, greater transparency in how records are collected, and expanded opportunities for scientific research on linking records and survey data. But so far no existing research has tested this approach; future research should examine this idea to explore any promise that it may hold.

Areas for future research:

- Identifying the ways that linkage errors influence subsequent statistical analyses
- Identifying methods to assess the accuracy of administrative records
- Best practices for selecting the most appropriate linkage method
- Effects of consent rates and consent bias on the quality of linked records
- Methods for optimally linking three or more datasets
- Methods for making better use of administrative records prior to survey data collection
- Improving statistical matching algorithms to overcome low linkage consent rates
- Incentivizing interviewers to obtain higher linkage consent rates
- Identifying optimal approaches to using multiple linkage techniques in sequence

Linking Knowledge Networks Web Panel Data with External Data – Josh Pasek

In addition to official administrative records, such as those collected by government agencies, medical facilities, or employers, another class of records data also exists in the form of consumer file data collected by market research organizations. These data can be purchased from clearinghouses that aggregate the data and make it available to marketers and market researchers. InfoUSA, one of the largest data clearinghouses, has a database of over 3.5 billion records from over 75 sources to which they sell access. Other sources of these large datasets include Experian,

Acxiom, and Marketing Systems Group. These databases typically contain individual-level demographics, addresses, and other general information.

Survey researchers are interested in these data because, to the extent that they can be linked to the individuals selected using traditional survey sampling procedures, it may be possible to develop a better understanding of not only respondents but also non-respondents. There are several potential benefits that linking consumer databases with survey sampling frames could confer. First, it could help improve the efficiency of sampling; by integrating the demographic and other ancillary data with the sampling frame information, it may be easier to accurately oversample hard-to-reach populations. This could lead to lower costs of data collection and more representative samples of the population being surveyed, which might reduce bias and the need for large *post-hoc* survey weights. Another potential benefit from these data are that they can provide substantially more information about non-respondents than is commonly available, which has implications for how nonresponse bias calculations are made and might even enable researchers to correct for biases in the sampling frame. Taken together, the potential benefits for using these data in conjunction with surveys is difficult to overstate.

However, all of these potential benefits are predicated on the assumption that the data in the consumer databases are of sufficient quality to serve these purposes. The two primary dimensions of quality that need to be considered are accuracy and completeness. If the consumer file data are inaccurate, then linking with survey data could have deleterious rather than beneficial effects because it could lead to incorrect inferences or might lead us to misstate the nature of nonresponse biases. If the consumer data are incomplete, then there could be differential levels of certainty in the quality of the match with the survey data, which adds another layer of complexity to the already challenging problem of missing data. These are important questions that need to be addressed by future research.

The current state of research seems to indicate that market research databases are not yet of sufficient quality to justify their widespread use in conjunction with survey data. The consumer file data are not particularly accurate, estimates derived from these data frequently differ from self-reports, missing consumer data is systematic and non-ignorable, and standard imputation algorithms do not fully correct these biases. However, the potential benefits from these databases certainly warrant considerable future research into how these sources of information can best be used in valid ways. Research is needed identify appropriate and valid uses for these data along with novel ways for correcting the biases and inaccuracies in the consumer marketing data.

One important feature of these market research databases that severely limits their utility for serious research is the lack of transparency with regard to how the data are collected, matched internally, and manipulated prior to bundling for sale. The companies that sell access to these data nearly universally claim that this is proprietary “trade-secret” information, and they refuse to give access to these key features of the data that would enable researchers to evaluate their quality. Until these practices change, it is unlikely that these databases will be able to provide their maximum benefit to researchers. Serious researchers then should restrict their use of these commercial data to experiments and exploratory analyses; applications for substantive analyses should be avoided until greater research transparency can be achieved.

Areas for future research:

- Identifying organizations with commercial databases that are willing to engage in transparent research practices
    - When the records were obtained
    - From whom the records were obtained
    - How the records are cleaned
    - Full disclosure about the matching algorithms, including the criteria for determining a match
- Assessing the correspondence of data from consumer databases and survey self-reports
- Evaluating the nature of missing data in the consumer databases
- Exploring approaches to correct for missing data in consumer databases
- Using consumer data to examine variables that might not be collected by surveys as a supplemental form of data and conducting sensitivity analyses on these data
- Determining whether consumer file data identifies differences between respondents and nonrespondents that can be used to improve survey weights.

Linking Survey Data with the Catalist Commercial Database – Bob Blaemire

As a political corollary for the consumer marketing databases that have been developed for market research, politically oriented organizations have begun to aggregate data about individuals as well. Rather than simply collecting demographic or purchasing behavior data, these organizations attempt to identify people that may be politically active or amenable to becoming politically active given their past behavior. The types of records that these organizations collect are membership rosters for other groups or clubs such as the Sierra Club in the case of Democratic organizations or the National Rifle Association for Republican organizations.

The application that these organizations typically have in mind for such political databases is linking them with official voter registration and turnout records that are made publicly available by each state for non-commercial purposes. By combining these datasets, the notion is that political parties and politicians may be able to micro-target individuals that may have a higher probability of voting for some particular candidate or issue if they were contacted. These databases can be nearly as large as those generated by market research firms; for example, Catalist, the leading Democratic-leaning database creator has records for over 280 million individuals in their database.

For many of the individuals represented in these databases, there are extensive records of past voting behavior and group membership. So instead of mailing an identical political advertisement to every member of a state's Democratic party, the database organization can subset down to the individuals that perhaps haven't voted consistently in prior elections but who may be easily persuaded to turn out to vote. Going a step further, these organizations can then use the past group membership history for each individual to create subgroups that will receive different political advertisements based on what the organization believes will constitute a

persuasive argument. Beyond political advertising, these resources have been used to assist with political fundraising, volunteer solicitation, and get-out-the-vote campaigns.

Recently, there has been considerable interest on the part of some researchers that make use of survey data for political applications, such as political scientists and polling organizations, to attempt to use these databases. Often, the goal is to validate turnout using a one-stop-shop approach rather than having to conduct the extremely varied and often challenging task of collecting voter files from each state after an election. In theory, using a political database organization to do this proverbial "heavy lifting" could make validation studies much easier and give survey researchers and pollsters a better sense for how to handle self-report data regarding voting behavior, which is widely believed to be over-reported.

However, there are significant concerns that must be addressed by those using these sorts of databases for academic research. First, they are constructed very similarly to the commercial consumer marketing databases and often even incorporate some of the demographic data purchased from these databases. This means that they may suffer from some of the same data quality issues associated with the consumer databases. Second, the political databases frequently have to make decisions about conflicting records for the same individual; rather than keeping all of the records, they use a statistical model to determine which record should be kept and then delete the conflicting records. Lack of transparency in this data cleaning process precludes a very important step necessary for serious researchers to evaluate the quality of the data in the database. Third, these databases are designed with a focus on coverage, the goal being to increase the number of people that they can categorize in their models. This may come at the cost of accuracy if many records with very poor-quality data are aggregated with those records that may actually be of sufficient quality for serious academic or government research in such a way that the researcher cannot disentangle the two. Lastly, the transparency and rigor in documentation of procedures and changes to the data may be below ideal standards for academic or government research or may be held as proprietary information and not made available to researchers. It is also true, however, that academics could know a lot more about the people they may be contacting than they do by using anonymous surveys. The potential utility for researchers in these databases is certainly great, but much of it will likely go unrealized in the wider academic and government audiences until database organizations are able to provide greater transparency into their methods and allow researchers unencumbered access to evaluate data quality issues.

Areas for future research:

- Increasing transparency into the sources and quality of the data in private commercial or political databases
- Establishing and making public metrics for data quality for each private database
- Identifying the sources of errors and approaches to correcting these errors in private databases
- Developing testing methods to identify the error rates in private databases

Election Administration Data – Michael P. McDonald

Historically, survey respondents' self-reports of voting have exceeded official records of turnout by substantial margins. This led to longstanding skepticism about respondents' reports of their turnout behavior. Some suggest that respondents may think it's embarrassing to admit not voting and therefore claim to have voted when they didn't. Others propose that the over-reports may be due to people who usually vote in every election but happened to overlook the fact that they uncharacteristically didn't vote in a particular instance. Because the official records are a gold standard, many researchers have argued in favor of replacing the self-report data with the official government records.

There are two general types of voter files that are frequently and reliably compiled into commercial and political databases, voter registration and absentee ballot requests. Vote history is a common component of these data as well, but their reliability is variable and access to the general public can be restricted. These voter files are used for a variety of research purposes ranging from vote validation studies and voter mobilization studies to racial bloc voting analyses for voting rights litigation.

Because of the gold standard status of voter files, they hold considerable promise, in theory, for linking with survey data that include variables relating to any of the data contained in these files. However, these files are not without error. The ANES, after a long history of using voter files to conduct turnout validation studies, determined that the inaccuracies and errors were sufficiently egregious and the task of validation sufficiently challenging that they stopped conducting the validation studies altogether.

The challenges that the ANES encountered were not isolated. For voter registration alone, there are common errors in the records themselves and issues with match algorithms across databases often leads to false positive and false negative matches, such that a person who did not vote may be identified as a voter or a person who did vote may be identified as a nonvoter. Furthermore, there is significant variability in how states handle voter files. Some states will release voter files only to political campaigns, meaning that they are unavailable to academic or government researchers. In the case of Florida, even federal programs are unable to get access to the absentee files. Similarly, Virginia forbids disseminating or sharing the vote history data outside of political campaigns. Sometimes these files can be exceedingly expensive to purchase outright or too time-consuming to collect, especially when vote history is available only from local election officials.

These complexities are part of the allure of using a pre-packaged database created by a commercial vendor such as Catalist. Allowing the organization to collect and manage all of the data can save researchers considerable time and effort. However, this outsourcing comes at the cost of the data becoming a 'black box' with very little information about their quality or the validity of using them for rigorous research. Commercial vendors' business plans can be at odds with research agendas. Vendors want to provide an accurate current snapshot of voter registration to their clients for campaign purposes. The most current voter files available from election officials may exclude individuals that are have become ineligible since the last election, most often because they moved. The most current voter file may thus not contain records of all voters who participated in a given election. Commercial vendors may perform their own purging of records or enhance the data in other ways, such as attempting to match registered voters who

moved with the record in their former home. Little is known about the potential biases that may arise from these practices. These potential errors are important to understand, as a recent study claiming widespread noncitizen voting exemplifies: are five vote-validated self-reported noncitizens indicative of true rates of noncitizen voting, or are they artifacts of commercial vendors' data processing techniques?

Areas for future research:

- Identifying better methods for collecting and aggregating voter file data from states
- Working with database vendors and data clearinghouses to identify data files that are of sufficient quality for academic and government research
- Understanding better potential errors in database management procedures resulting from election administrators and commercial vendors practices

Challenges with Validating Survey Data – Matthew K. Berent

Self-reports of voting behavior in many studies, including the ANES, are often about 20 percent higher than the official statistics on turnout. This has been a remarkably consistent finding since the 1960s, meaning that as actual turnout increases or decreases, there are similar increases or decreases in self-reports of voting. Many people take this information to imply that the estimates generated by self-report survey data are not representative of the population because they don't match the 'true' gold standard value provided by the government.

However, there are a number of different ways that this discrepancy could arise that warrant consideration. For example, a respondent might not answer the question because perhaps they believe voting behavior is private and sensitive and are unwilling to report their answer to the survey question; this leads to nonresponse and inaccuracy in the measure that is unrelated to the representativeness of the survey sample.

The second explanation is that there could be survey effects, meaning that participating in the survey could influence the voting behavior that the survey is attempting to measure. Many voting behavior studies involve very in-depth interviews before the election and then an interview after the election. It is reasonable to expect that, after having answered a long battery of questions about politics before an election, respondents are made aware of their political attitudes that were perhaps not as salient before the interview. This increased salience could then influence some respondents to go vote when they might not have otherwise. The result is that these surveys could inadvertently be biasing their own sample to over-represent voters.

This raises a third potential mechanism for higher levels of voting behavior in surveys than in the general population. It is possible that survey response propensity is positively associated with the propensity to vote, meaning that the same types of people who are willing to respond to a long interview about politics may be the same types of people who go out and vote. This would be another way that the sample of survey respondents could be biased from the general population, not demographically but behaviorally.

Lastly, it is entirely possible that respondents misreport their voting behavior, meaning that their answers are factually incorrect. This could be caused by respondents misinterpreting the question, misremembering the behavior, reporting based on their typical voting behavior rather than their specific behavior in the focal election, or intentional lying, which might be due to social desirability bias if the respondent believes that it would reflect poorly on them to report not voting in the election to the interviewer.

Misreporting is the most commonly cited cause of the discrepancy between survey data and the official records, leading to the typical conclusion that official records should be used instead of self-reports rather than attempting to identify the actual mechanism behind the discrepancy. However, the approach of using official records is not completely straightforward. To employ this method, researcher must obtain the official records of turnout histories, match each respondent to his or her official record, and then determine the "correct" turnout status for each respondent.

There are two primary problems with the validation task. The first is that over 200 million people are currently eligible to vote in U.S. elections; thus the matching task is not trivial simply on the basis of the size of the databases involved. Second, the federal government does not aggregate individual voting records, meaning that researchers must collect the records from each individual state, and there is substantial variability between states in the difficulty of obtaining these records.

Researchers are increasingly turning to commercial vendors of political and voter file databases to conduct the matching between survey data and the official voter records. However, these researchers face another set of challenges when working with these databases. For example, vendors vary in their level of transparency, which means that researchers need to identify and specify the level of uncertainty that they are willing to accept with regard to the provenance and quality of their data. Vendors are also typically unwilling to provide complete details about their matching algorithms, which makes it impossible to estimate the reliability or validity of the matches.

Some researchers opt to attempt in-house matching by obtaining the government registration and turnout records from a sample of states. Publicly available computer applications, such as LinkPlus from the Centers for Disease Control, can then be used to match survey respondents to their official government records. This approach requires more work on the part of the researcher, but the benefits are complete control over the data cleaning and matching processes.

Recent evidence from an ANES in-house matching project indicates that two main factors are contributing to the discrepancy between self-reports and official records of voting behavior. The first is a downward bias in government records that occurs when the records incorrectly identify some respondents as having not turned out when in fact they did. These are all cases where a record match cannot be found, not cases where the self-report and government record data disagree explicitly. The second factor is an upward bias in the self-reports that occurs because people who participate in surveys are more likely to vote as was discussed above. These biases are additive, not offsetting, and seem to account for nearly all of the discrepancy between the self-report data and government records.

Areas for future research:

- Developing a better set of tools for researchers to conduct their own transparent validation studies without needing to use commercial vendors
- Identifying and understanding the correlates of survey participation and turnout to better understand the potential mechanism that drives the higher levels of turnout among survey respondents

Improving Government, Academic and Industry Data-Sharing Opportunities – Robert Groves

The social sciences seem to be at a key point in their history. Part of the apparent transformation is that research that has traditionally been the purview of academia and government is increasingly being done in the private sector. In the past, the social sciences, and survey research in particular, operated under a model in which researchers would create data, analyze them, and then make them publicly available for use, usually because the research had been federally funded. This paradigm seems to be shifting to more of a market-based approach where massive amounts of data are collected about individuals using both traditional social science methods and new methods that are often focused on reducing data collection costs. Meanwhile, the costs associated with generating the traditional high-quality social science data, such as those from nationally representative surveys, are at an all-time high.

Some of the most important features of government and academic produced survey data are that they are incredibly multivariate, they provide uniform and consistent measurement, and researchers control them. The researchers identify a specific construct that they want to measure and then invent the tools to measure these constructs, and when this process is done well these data can provide rich insights into the social world. Deep care is exercised in constructing each measurement, assuring that it reflects accurately the underlying construct. Further, many measurements are taken on the same unit, so that multivariate models can be estimated from the data. Finally, the measurements are administered in a consistent manner, achieving comparability over units.

Another benefit of these data is that they are typically made publicly available so that other researchers can use them an unlimited number of times. Sociology would not be the same field it is today without the General Social Survey; the same is true of political science with the American National Election Studies and economics with the Panel Survey of Income Dynamics. These specific studies provide the additional benefit of representing a long time-series of data that allow longitudinal trends to be easily examined in consistent and valid ways. In most of the social sciences, generations of quantitative scientists have “cut their teeth” on these data sets, and thousands of scholarly articles and books document the information extracted from these data. They have yielded important discoveries about the society and its economy.

However, the new data being generated in the commercial sector are often not nearly as rich or powerful as the survey data of the past. These data are much more likely to be univariate, lacking in consistency and uniformity, and “organic” in the sense that they are not controlled by researchers. Private sector firms disproportionately hold these data, and these firms often have no

chief mission to benefiting society. There is nothing in their goals supporting making these data freely available for government or academic research purposes. Rather, for an increasing number of these firms, these data are viewed as a potential source of revenue to be guarded carefully and kept proprietary.

At the same time as commercial firms are collecting a disproportionate amount of data, academic and government research costs have become unsustainable, with no evident solution in sight. Research costs have risen, especially for those methods that rely on the general public to supply self-reports in the data collection. In most developed countries, public participation among sampled persons is declining, forcing data collection agents to increase the number of repeated attempts to gain their participation. At the same time, the quantitative side of the social sciences seems to have won the day, with nearly every institution in society now primarily using quantitative information. Taken together with the unsustainable costs of traditional research methods, this raises the specter of privatized and commercialized social science, at least to the extent that firms can profit from statistical information about the population.

“Big data” is a buzzword that is commonly used across research sectors, and nearly everyone agrees there is unexplored potential value in these vast stores of data. However, there is little agreement about how government or academic researchers should acquire and process these data, and at a more fundamental social science level, identifying appropriate inferential frameworks for these data is also contentious. Further complicating matters, commercial firms that hold much of these data are hesitant to make them available to government and academic researchers for a few reasons that need to be addressed by future work.

First, these companies are concerned about liability if their data are used for linking with government records and then somehow breached. This will remain a concern for these firms until legislation protecting them is put into place, so it will be important for a consortium of commercial, government, and academic organizations and individuals to attempt to make headway on these legislative efforts.

Second, these firms are concerned about increased attention to confidentiality because many of their data collection models depend on much of the population not paying attention to or caring about the confidentiality of their data. This concern needs more attention from the perspective of government and academic researchers who have typically approached confidentiality from the exact opposite position, often even going so far as to make confidentiality an argument for participation. Until common ground on data confidentiality can be addressed, commercial data are likely to remain sparsely available.

Lastly, commercial firms are concerned that the use of their data may lead to a potentially profitable product being generated on which they will be unable to make money because of the collaborative agreement. This is a concern that government and academic researchers will have a harder time addressing but one that needs to remain at the forefront of thought as these joint data use efforts are arranged. Thus, an important area in need of future research and funding is bringing government, academic, and commercial data interests together. This is a popular topic of conversation among social scientists but one that has not translated well into large-scale fundable research programs.

The starting point for this work needs to be a data integration research program that brings together the highly related but traditionally disparate interests of computer science, statistics, mathematics, and the social sciences. This big data consortium also needs to be structured in such a way that it addresses the concerns of commercial firms and appeals to the interests of government statistical agencies. This will fundamentally be an administrative endeavor aimed at building the infrastructure that will enable science to progress while also benefiting government and commercial interests. This is an area in which NSF could directly and immediately spur progress.

Continued funding of piecemeal disciplinary projects purporting to make progress in this area is insufficient and is likely to restrict significant progress more than enable it. Interdisciplinary funding at a higher level than the traditional programs of political science or sociology is needed to make substantive progress in this area. Status quo funding that does not address the larger infrastructure needs will result in continued reinvention of linkage, estimation, and scrutiny techniques between disciplines that are unaware of the progress being made in each other. This endeavor is considerably larger than any one discipline and as such requires the development of a cohesive complementary set of research programs that are both interdisciplinary and cross-sector. Funding agencies like the NSF need to recognize the long-term benefits that these efforts could garner and make concerted efforts to incentivize work in this area.

Section 4: Improving Research Transparency and Data Dissemination

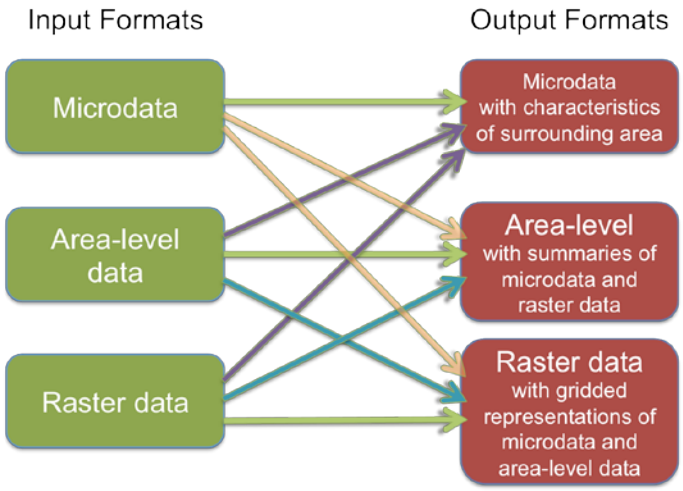
Data Curation – Steven Ruggles

Data curation is often overlooked when individual researchers or longstanding studies apply for new or continuing funding. Many researchers are more concerned about maximizing funds that can be applied toward collecting new data rather than appropriately integrating, disseminating, and preserving existing data. With new data collection costs at all-time highs, it may be time for researchers to begin investing in maximizing the utility of existing datasets. There are four major data curation challenges: (1) Data Integration, allowing interoperability across time and across data sources; (2) Electronic Dissemination, including on-line tools for data discovery and manipulation, (3) Sustainability, including planning for long-run preservation and access, and the creation of persistent identifiers, and (4) Metadata, machine-processable structured documentation.

Data integration is important because it enables time-series analyses to be undertaken without each individual researcher generating their own systems for harmonizing differences in datasets. Longstanding studies evolve over time, modifying their survey instruments, data collection procedures, processing protocols, and archiving methods. These factors make cross-temporal analyses challenging. Researchers are forced to adopt *ad hoc* data integration solutions, which leads to inconsistencies between researchers. Even modest investments in data integration can reduce redundant effort and minimize the potential for introducing error.

For example, in 1991 microdata (individual-level data) samples from the Decennial Census existed for ten census years dating back to 1850. These data had ten different codebooks associated with them totaling 2880 pages of codes and over 1000 pages of ancillary documentation. Furthermore, of these ten codebooks, nine of them used different coding systems for most variables. The Integrated Public Use Microdata Series (IPUMS) project harmonized the codes, generated a consistent record layout, and integrated the documentation, all with no loss of information. In 1999, IPUMS was expanded to include 100 national statistical agencies around the world. The result is that over 500 censuses and surveys have been harmonized for 75 countries and nearly one billion person-records, and this is set to double over the next five years.

These approaches and best practices for harmonizing datasets are also being extended to data with different formats and from different scientific domains. For example, many of these individual-level datasets can be appended with community-level data relating to the physical environmental contexts in which individuals were and are living. This might include data about land-use statistics, land cover from satellite imagery, raster data (gridded values linked to spatial coordinates), or historical climate records. Data integration approaches like this can add richness to already fine-grained datasets by providing important contextual variables.



Some of these principles have been applied by surveys such as the ANES, GSS, and PSID to their datasets, meaning that researchers can access cumulative data files that have been harmonized. However, there is very little documentation of cross-temporal comparability or consistency issues. Future funding and research efforts on data integration can focus on designing surveys to maximize interoperability across time and between surveys by developing and implementing standard classifications and coding systems. This would increase the value and use of all types of survey data designed with these features.

The second major topic relating to data curation is dissemination. Efficient data dissemination is important because the large investment of scarce resources for social science infrastructure can be justified only if the data are widely used. Even modest investments in data dissemination can reduce redundant efforts and stimulate new kinds of research. Increased emphasis on ongoing improvements to dissemination methods and platforms is necessary to maximize the value of survey data.

Best practices for dissemination suggest that data storage and access platforms undergo constant improvements to keep pace with technological advancements and user expectations. For example, IPUMS is now onto its 4th major generation of dissemination tools; these are driven entirely by structured metadata stored in a relational database. The tools include the ability to manipulate the data easily, online analyses, and the use of Data Documentation Initiative (DDI) standards for metadata, which is discussed further below.

Most survey data dissemination platforms, including ANES, GSS, and PSID, do not follow these best practices. The platforms have not kept pace with technological advancements and are for the most part running on antiquated systems that do not allow efficient discovery or exploration of variables. Furthermore, the metadata for these surveys are typically in PDF documents, effectively locked away and not actionable, which severely hinders the efficiency of any variable identification or analysis tools. The following are some key statistics for the metadata of the "Big 3" surveys:

- ANES
  - ~245 PDF files
  - ~35,000 non-actionable pages
- GSS
  - ~250 PDF files
  - ~25,000 non-actionable pages
- PSID
  - ~230 PDF files
  - ~50,000 non-actionable pages

To begin making these data more available and useful, the ANES, GSS, and PSID should implement metadata browsing functionality that will enable researchers to quickly identify variable availability across time and to directly access survey instruments and supporting documentation. Better dissemination software for selecting subsets of variables is also important, especially for PSID. These surveys also need to implement machine-understandable metadata; even in the form of Stata, SAS, and SPSS setup files, the current approaches tend to be inadequate or nonexistent. Future funding and research to begin optimizing the usefulness of existing survey data is critically important.

The third key point about data curation relates to sustainability. Funding agencies, universities, and government statistical agencies have invested hundreds of millions of dollars in data infrastructure. The greatest value from this infrastructure and the data generated by it comes from the power for cross-temporal analysis. In general, the older the data, the more rare and irreplaceable it is – and the greater value it has to researchers. This suggests that researchers, survey organizations, and sources of research funding must all work together to ensure effective stewardship of the nation's statistical heritage.

Broadly speaking, sustainability with regard to data curation means preserving data, especially old data. This preservation must also include steps to maintain broad access to the data. Sustainability can be thought of in three key areas: organizational, financial, and technical

sustainability that will enable the data to be migrated effectively as technology changes. The key question that ongoing surveys need to address is whether or not they are functioning as effective data archives. If they are, then they need to implement a formal preservation plan, and if they are not, then they need to arrange a cooperative agreement with a major archive such as the Inter-University Consortium for Political and Social Research (ICPSR). Current integration with ICPSR is varied across the major NSF surveys: ANES share the most with 57 of 68 data files being in ICPSR, GSS is next with 16 of 36 files, and the PSID has only 4 of about 100 files in ICPSR.

To move toward a sustainable data curation model, surveys also need to expand the use of persistent identifiers. In the current state of affairs, identifying information for variables, datasets, and documentation are prone to changing, and in many cases have changed multiple times and exist in multiple iterations. Survey organizations need to adopt consistent standards for unique identifiers that will be consistent over time. The NSF could speed this transition by requiring that funded datasets provide a persistent identifier such as a Digital Object Identifier (DOI) or another recognized persistent identifier and present a plan for long-term maintenance of the DOIs.

The fourth and final topic for data curation is metadata. As has been mentioned elsewhere in this report, metadata is a class of information that relates to and describes that focal data – it is "data about data". Examples of metadata include technical reports, survey instruments, interviewer instructions, show cards, and other documentation of the survey process or variables. The key point for survey organizations is that machine-actionable structured metadata is vital to enabling efficient data integration and dissemination. Without good metadata, software must be custom-built for each dataset, which is terribly inefficient and not a good use of resources. Metadata following DDI standards are also necessary for preservation purposes. In short, all aspects of data curation require good metadata in order to function properly. Currently all major survey data archives use DDI including: ICPSR, UK Data Archive, Odum Institute, and the Institute for Quantitative Social Science. Unfortunately, the ANES, GSS, and PSID likely have over 100,000 pages of documentation that need to be converted to a structured metadata format, which means that significant investment and future work is required. Future research aimed at reducing the costs of conversion through the development of "smart" automated processes could have great impact on this process.

Areas for future research:

- Improving the state of metadata held by the ANES, GSS, and PSID
    - Identifying a set of common standards for generating, structuring, and disseminating new metadata that conform with DDI standards
    - Identifying projects that might reduce the conversion costs for existing metadata
- Improving data integration within and between surveys
- Establishing standards for dissemination across surveys

Evaluating the Usability of Survey Project Websites – David L. Vannette

Website usability is not a subject that is unique to survey research; however, the principles and best practices for making websites usable have often been ignored or poorly implemented by

survey organizations. There are two primary reasons that funding agencies, survey organizations, and researchers should care about improving survey website usability. First, and most importantly, is dissemination; by making survey data and documentation easy to access through survey websites, we can increase the use of survey data and broaden the influence and importance of survey research. Second, improving survey website usability is important because it will make secondary research easier to conduct because both data and documentation will be easily located, accessed, and used. This will make the process of using survey data more accessible and transparent.

Broadly speaking, website usability means creating the conditions for an efficient and effective visit to the website in question. This means satisfying a number of important criteria that define what usability means in terms of broad goals, including:

- Providing relevant and easily accessible information
- Enabling learnable routines
- Designing efficient paths
- Creating memorable patterns
- Minimizing user errors
- Satisfying user experience

Breaking each of these items down further; in terms of providing relevant and easy accessible information to the users, it's important that those things go together. Information on survey websites should be both relevant to user needs and easy to access. Easy access implies using standardized formats and machine-actionable documentation and data so that users do not need to search through thousands of pages of PDF documents to find a single piece of information.

Next, it is key for survey website structure and organization to enable learnable routines. This refers to having similar tasks on a website follow similar routines so that users can apply learning from the browsing or searching that they did for data to the browsing or searching that they do for documentation. Similarly, survey websites need to design efficient paths, and this refers to minimizing the number of clicks or the number of search terms that users need to use before they find what they came to the website for. The key is to minimize the distance between when users arrive at the site and when they get to the data or documentation that they came looking for. Also in this vein is the best practice of creating memorable patterns in the ways that webpages are structured. Important elements should not move around on web pages or disappear on some pages and come back on others.

Lastly, in terms of general goals, survey websites should also seek to minimize user errors. An important goal is to make it really hard for users to make mistakes. Users should not get 10 minutes into a task and realize that they should have caught a mistake that they made when they first arrived at the website. This means making survey websites clear and intuitive for the average user. Taken together, these broad goals are aimed at helping users to have a satisfying experience.

A very small proportion of all websites satisfy the goals outlined above, but survey websites can be particularly egregious in not achieving some of these usability goals. A good starting point for assessing website usability is asking what their users want, and in the case of survey websites this is thankfully not a challenging task. Survey website users are typically looking for three broad categories of information: 1) data, 2) process and data documentation, and 3) examples of how these data are used such as publications.

After identifying the goals of website usability and the particular needs of survey website users, the next step is to identify best practices that can guide the design decisions implemented by survey websites. There are many best practices that have been defined for website usability, but they fall into two broad categories of principles: 1) optimizing for memory and 2) optimizing for visual perception.

In terms of optimizing for memory, it is important that websites standardize task sequences. This means that finding variables and documentation on a survey website should follow the same procedure across all possible pages and tasks. Standardized task sequences allow users to apply what they learned about conducting a task in one part of the website to all other parts of the website where that task is performed. At the highest level, there are two classes of tasks that users can engage in on a website: 1) searching and 2) browsing. Survey websites need to ensure that no matter what part of the site the user is in, these two tasks are structured the same. This means that once a user knows the task sequence for searching for data, they also automatically know the task sequence for searching for documentation. This is an area in which survey websites such as the ANES, GSS, and PSID need considerable development and improvement.

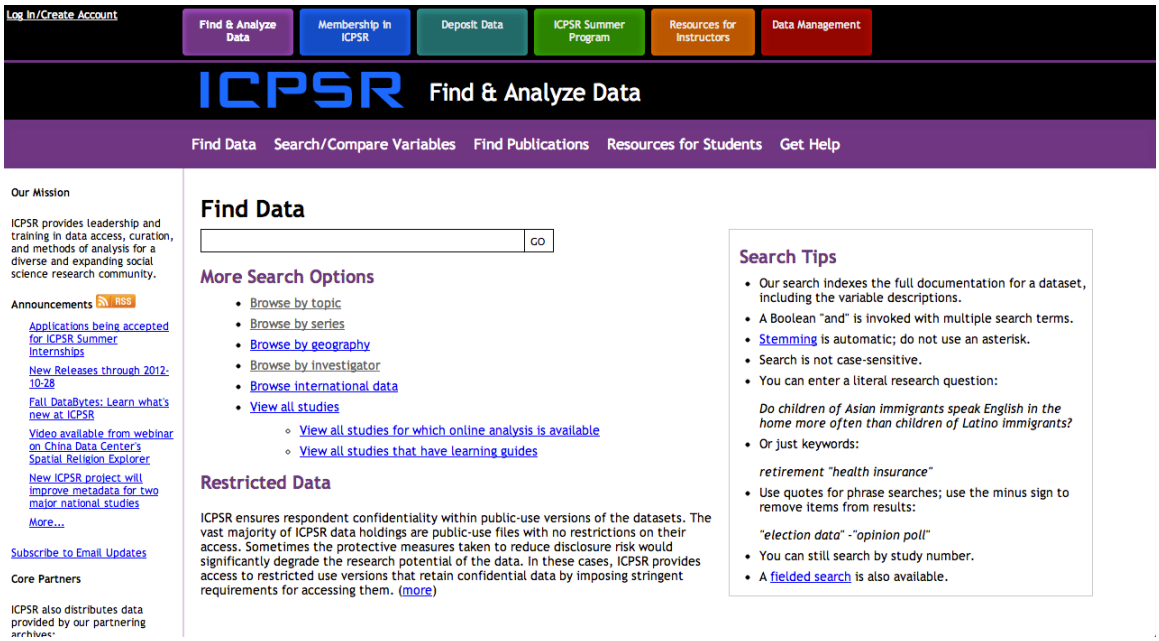
Reducing user workload and designing websites for working memory limitations are also key best practices. For example, the GSS website uses a system called Nesstar for online data retrieval and analysis. This system requires knowledge of a 51 page user guide, and this knowledge is not transferable because it is specific to the GSS and even there only to the Nesstar portion of the website. This design approach does not minimize user workload or acknowledge the working memory limitations that new users have when coming to the site. Survey websites should take care to display directly usable information to users. For example, on the starting pages for datasets, important survey design and dataset features should be prominently displayed so that users can quickly evaluate whether this is indeed the dataset that they wanted to access. The websites for the ANES, GSS, and PSID can all improve in applying these best practices.

Survey websites often need to house tens of thousands of pages of technical documentation and code books. However, this information is most commonly archived in PDF documents that are not machine actionable, which in a very basic sense means that a website search does not include searches inside these documents. These documents are also poorly linked and can be nearly impossible to use without extensive prior knowledge of the particulars of the documentation. Best practices for survey documentation suggest using organizing features such as linked tables of contents and structured documents with labeled sections. Tying in with the importance of accessibility with regard to data curation, it is also important for all documentation to adhere to widely accepted standards such as DOI and DDI (see report section on Data Curation). These are practices that have not been adopted by survey websites such as ANES, GSS, and PSID.

A number of other best practices are well exemplified by the ICPSR website. These best practices include:

- Optimizing information density
- Aligning elements consistently
- Using fluid layouts that maximize screen size and resolution
- Helping users search in addition to browse

These design features, shown in the example from the ICPSR website below, lead to a much more user-friendly experience relative to other survey websites.



One of the biggest challenge facing survey organizations is determining how to allocate scarce resources. Survey staff often view every dollar spent on something other than data collection as money that could have been spent better. This perspective can be seen in the current state of the websites of the ANES, GSS, and PSID, where usability, data curation, and transparency best practices are often ignored or poorly implemented. Survey website usability requires increased investment and attention in order to optimize all of these best practices. If survey organizations are not able to implement these best practices on their own, then they should partner with other archives such as ICPSR to ensure that their data are made more available and usable to a broader array of potential users.

Lastly, best practices dictate conducting usability studies when making changes to websites. The current state of many survey websites seems to imply that little if any systematic usability research was conducted. Even small-scale studies can indicate important problems or areas where improvements could easily be made. On a similar note, these survey organizations have considerable expertise in survey data collection, and it might be useful for them to consider

applying that expertise toward identifying what their data users like and dislike about the current websites and features that could be implemented to improve the user experience.

In summary, there are many best practices for improving website usability, but many of these practices and principles are not implemented by survey websites. This presents an opportunity for future work by these survey websites on improving their structure and design. Improving website usability also provides a platform with which survey organizations can display other best practices in the areas of data curation and transparency.

Areas for future research:

- Identifying practical approaches for the websites of the ANES, GSS, and PSID to increase their usability
- Conducting usability research on existing websites to identify easily implemented changes to improve usability
- Conducting surveys to identify what features users find helpful or problematic

Research Transparency and the Credibility of Survey-Based Social Science - Arthur Lupia

Many scientific researchers are motivated to produce credible scientific contributions. At the same time, many face pressures to produce and publish findings as quickly as possible. With such pressures comes the temptation to ignore critical assumptions and to view rigorous documentation and reporting as secondary tasks to be completed later. I contend that limited introspection and documentation on the part of researchers threatens the credibility and legitimacy of scientific research as a whole.

Survey research is implicated in such problems. Every survey-based claim follows from arguments whose conclusions depend on the truth-values of important assumptions. Some of these assumptions are statistical and others relate to the design characteristics of the study. Limited introspection about these key methodological assumptions puts scholars at risk of promulgating false claims.

Indeed, in several disciplines, scholars improperly interpret or have been unable to replicate highly visible survey-based claims. Collectively, these failures undermine the aggregate credibility of survey research. There are a number of constructive steps that researchers and organizations can take to increase the credibility of many survey-based endeavors. Such efforts should focus on ways to improve research transparency, meaning documenting the processes, decisions, and actions that convert labor and capital into survey data and then into survey-based knowledge claims.

Research Transparency has two components: production transparency and analytic transparency. Production Transparency implies providing information about the procedures that convert labor and capital into data points. Today, many surveys provide little to no information about these procedures, which include but are not limited to: sample selection, respondent recruitment, question selection, question wording, response options, directions to interviewers, post-interview processing, and coding of open-ended responses. Many users, in turn, passively accept survey

data as accurate rather than a product of practices that can produce biases and inaccuracies if misunderstood. When users cannot access information about how survey data was produced, their ability to make accurate survey-based claims can be severely inhibited.

Analytic Transparency implies a full account of how a researcher drew conclusions from a given survey, clearly mapping the path from the data to specific empirical claims. There are many examples of published work in which such paths are neither public nor reproducible. By meeting a small number of transparency requirements at production and analytic stages, scholars can protect their own credibility and contribute to the credibility and perceived legitimacy of survey research more generally.

Analytic transparency means making every step in a research process, including errors, open to evaluation. Methods courses across the sciences emphasize the value of documenting research processes in sufficient detail that an outside researcher could replicate the finding. Beyond simple replicability, our goal is to provide fellow researchers and outside onlookers all possible information available to evaluate the value of the contribution of the research, not just the information that leads to judgment in one particular direction. In terms of best practices, this implies that researchers should consider relying on a system of "lab books" that document all decisions made regarding a research project and are then made available when the project is published.

Effective data citation and curation practices are critical components of efforts to increase the availability of production and analytic materials and to increase incentives for transparency. If data and documentation are not made readily available to users in easily accessible formats, the benefits of increased transparency are reduced. Survey organizations should document all decisions that have any influence on data or data quality and make this documentation easily accessible online using best practices for data curation and website usability.

Funding agencies such as the NSF and academic journals can also play important roles in increasing research transparency. By developing and implementing data sharing and documentation transparency requirements, funding agencies can boost incentives for transparency in many different kinds of research. Journals, in turn, can require authors to deposit data and documentation to trusted digital repositories as a condition of publication. If funders and journals followed a similar model, it is likely that scholars across many disciplines would adopt new transparency norms and practices, which would benefit all of the social sciences.

Public trust in, and the public value of, scientific research depends on the extent to which scientists are willing and able to share the procedures by which their knowledge claims were produced. Scientific integrity of this kind corresponds to a kind of utter honesty, bending over backwards to make certain that evaluations of the work are made with complete information. For scientific researchers, the sources of our distinctive legitimacy are rigor and transparency. The elevated social value of scientific research comes from the ability of others to evaluate the meaning of our findings. Increasing transparency is a critical step to the continued production of social science in the public interest.

Areas for future research:

- Identification and publication of a set of transparency standards for social science research funded by NSF
- Creation and implementation of institutional means to support greater transparency
- An NSF data and documentation registry and repository for all NSF funded research

Conference website and organizer contact information

Conference website: https://iriss.stanford.edu/content/future-survey-research-nsf

Conference Conveners	Affiliation	Contact
Jon A. Krosnick	Stanford University	krosnick@stanford.edu
Stanley Presser	University of Maryland	stanleyp@umd.edu
Kaye Husbands Fealing	University of Minnesota	khf@umn.edu
Steven Ruggles	University of Minnesota	Ruggles@umn.edu

Conference Organizer	Affiliation	Contact
David L. Vannette	Stanford University	vannette@stanford.edu

List of Conference Presenters

Presenter	Affiliation
Kathleen T. Ashenfelter	U.S. Census Bureau
Robert Belli	University of Nebraska-Lincoln
Matthew K. Berent	Matt Berent Consulting
Robert Blaemire	Catalist
J. Michael Brick	Westat
Curtiss Cobb	GfK-Knowledge Networks
Matthew DeBell	Stanford University
Robert Groves	Georgetown University
Scott Keeter	Pew Research Center
Stephen Kosslyn	Stanford University
Frauke Kreuter	University of Maryland
Jon A. Krosnick	Stanford University
Gary Langer	Langer Research Associates
Mark Liberman	University of Pennsylvania
Michael W. Link	The Nielsen Company
Arthur Lupia	University of Michigan
Michael P. McDonald	George Mason University
Colm O'Muircheartaigh	University of Chicago
Randall J. Olsen	Ohio State University
Steven Ruggles	University of Minnesota
Josh Pasek	University of Michigan
Joseph W. Sakshaug	University of Munich
Nora Cate Schaeffer	University of Wisconsin-Madison
Eleanor Singer	University of Michigan
Arthur A. Stone	Stony Brook University
Roger Tourangeau	Westat
David L. Vannette	Stanford University
David Weir	University of Michigan
Brady T. West	University of Michigan
Gordon Willis	National Cancer Institute

List of Conference Discussants

Discussant	Affiliation
Nancy Bates	U.S. Census Bureau
Paul Biemer	RTI
Johnny Blair	Abt SRBI
Norman Bradburn	NORC
David Cantor	Westat
Jon Cohen	Washington Post
Charles DiSogra	Abt SRBI
Jennifer Dykema	University of Wisconsin-Madison
Richard B. Freeman	Harvard University
Robin Gentry	Arbitron
Timothy Johnson	University of Chicago
Allan McCutcheon	University of Nebraska-Lincoln
Peter Miller	U.S. Census Bureau
Andy Peytchev	RTI
Trevor Tompson	AP-NORC

Conference Programs

The Future of Survey Research: Challenges and Opportunities

Sponsored by The National Science Foundation

Conference 1: October 3-4, 2012

Conference 2: November 8-9, 2012

Conference Conveners:
Jon Krosnick – *Stanford University*
Stanley Presser – *University of Maryland*
Kaye Husbands Fealing – *University of Minnesota*
Steven Ruggles – *University of Minnesota*

Conference Organizer:
David Vannette – *Stanford University*

Hosted at the Hilton Arlington
950 North Stafford Street
Arlington, Virginia 22203

STANFORD
UNIVERSITY

Conference Overview

The first portion of this two-part conference will focus on operational challenges and opportunities for survey researchers. The conference will begin with a new review of accumulated evidence about the widespread use of survey data and about the accuracy of those data. This evidence will set the tone for the conferences by making it clear that the method is both popular and reliable when implemented according to best practices.

The conference will then turn to elucidating the many challenges and opportunities facing survey researchers seeking to do so. Outlining those challenges and opportunities will entail illuminating current knowledge about how best to cope with them and maintain data quality while preserving affordable costs.

The net result of this conference will be (1) a set of insights about how surveys should be done today to maximize data quality (thereby specifying how major infrastructure surveys should be designed and carried out), (2) a list of the most important challenges and opportunities facing the methodology, and (3) a list of research questions that merit future investigation, perhaps with NSF grant support.

October 3rd Agenda

7:30-8:30AM ***Continental breakfast***

8:30-8:45AM **Reasons For Optimism About the Accuracy of Survey Research**
Jon Krosnick – Stanford University

8:45-9:40AM **The Impact of Survey Nonresponse on Survey Accuracy**
Scott Keeter – Pew Research Center

Among the many challenges facing survey research, the rise of non-response is perhaps the most worrisome. Nonresponse does not automatically lead to biased estimates, but greater nonresponse means greater potential for bias. Accordingly, a significant body of research has been devoted to assessing the impact of nonresponse on the accuracy of survey estimates. My presentation will review this research and provide a summary of what is known about the accuracy of surveys, especially those in the social and political realms where the NSF has been an important source of support for research. I will also identify areas in which additional research on nonresponse is needed.

9:40-9:55AM ***Coffee Break***

9:55-10:50AM **Proxy Reporting**
Curtiss Cobb - GfK

Many prominent studies, including the Current Population Survey (CPS) done by the U.S. Census Bureau, rely on proxy reporting. Specifically, the CPS accepts reports from any adult member of selected households, and these individuals are asked to describe many characteristics of all other household members. The accumulated literature on proxy reports suggests that although they may sometimes be quite accurate and comparable to self-reports, this is not always the case. Therefore, there appears to be value in understanding the actual accuracy of proxy reports, how to maximize their accuracy, and when to rely instead only on-self reports.

10:50-12:00PM **Coding Open Responses**
Arthur Lupia – University of Michigan (via video)
Dean Kotchka - Ascribe

Although survey researchers have been asking open-ended questions for decades and coding the obtained answers for later analysis, the procedures used for such coding have rarely if ever been informed by the large and multifaceted literature across the social sciences on best practices for content analysis of open-ended text. For example, it has not been routine for survey response coding to be subjected to assessments of reliability in light of the controversy in the content analysis literature about the merits of various different methods for doing such assessment. Furthermore, rarely have survey researchers released detailed sets of instructions used to guide their coders. With the rise in natural language processing by computers, the notion that the work of human coders may be better done by computers is gaining traction. These are just some of the potential opportunities for improvement in coding practices.

12:00-1:00PM ***Lunch Break***

1:00-2:00PM **What HLT Can Do For You (and vice versa)**
Mark Liberman – University of Pennsylvania

Some now-standard kinds of automatic text analysis, such as "document classification", "entity and relation tagging", and "sentiment analysis", are closely related to what social science researchers often do with transcripts of open-ended survey responses. Furthermore, automatic speech recognition can produce text that feeds effectively into analysis of this sort, often with relatively little degradation of performance. So in principle, "Human Language Technology" promises to give survey researchers faster results at lower costs. In addition, new sorts of insights may be available from analysis of the statistical patterns of word and concept associations in collections of survey transcripts, or from analysis of the many properties of spoken responses that are left out of the word sequences that make up such transcripts. And the growing scale and scope of networked social media raises the hope that "virtual surveys" can be done by appropriately-weighted analysis of the data available from such sources. But for the moment, satisfactory results will generally require the involvement of experts in natural language processing and speech technology, to select, adapt and apply the now-standard techniques, or to develop new ones. Luckily, there is an easy way to entice speech and language researchers into collaborations in this area. In this presentation, I'll give a basic taxonomy of relevant HLT methods, characterize their current level of performance, and describes the conditions needed for them to work well. I'll also describe the process that has led to steady technical progress in this field over the past 25 years, and explain how survey researchers can harness this process and direct it towards the goals that matter to them.

2:00-3:00PM **Address-based and List-based Sampling**
Colm O'Muircheartaigh – University of Chicago

In recent years, the process by which samples are drawn for face-to-face and mail surveys has increasingly shifted to address-based sampling via lists of addresses available in the USPS Delivery Sequence File, which is thought to be a near-universal and frequently updated frame of households. The efficiency offered by this sampling approach is appealing, but the specific procedures used to draw samples and to update listings in the field are not yet standardized across organizations.

3:00-3:15PM ***Coffee Break***

3:15-4:30PM **Interviewer Deviations From Scripts**
Nora-Cate Schaeffer – University of Wisconsin-Madison (via video)
Hector Santa Cruz – Stanford University

The dominant method in survey interviewing in recent decades has been standardized interviewing using instruments that are compatible with standardization. Standardization aims to make interviewers interchangeable to reduce the interviewer component of variance and thus increase the reliability of measurement. Practices of standardization have been confronted by at several challenges over the years. One is a concern with validity and comprehension and an interest by researchers in providing resources to interviewers to improve comprehension by respondents. A second is the development of interviewing methods in which data are requested and recorded in more variable forms that are less compatible with standardization, such as event history calendars. A third is evidence from the study of recordings that standardization is sometimes difficult for interviewers to maintain. Evaluating standardization and these new developments in interviewing is complicated by the difficulties of designing experiments that allow for adequate tests of the effects on both interviewer variance and validity or accuracy.

4:30-5:30PM **Panel Attrition**
Randall Olsen – Ohio State University

Some of the most important survey studies are long-term panels, including the National Longitudinal Surveys and the Panel Study of Income Dynamics. Part of their value comes from the tracking of the same individuals over very long time periods. But the longer the tracking period gets, the more sample attrition occurs, and the less representative of the population the sample may become. Although

some literature suggests that panel attribution has remarkably little impact on sample representativeness, other literature does not support this conclusion. Furthermore, efforts to prevent attrition are not standardized across studies and may not yet be optimized.

October 4th Agenda

7:30-8:30AM ***Continental breakfast***

8:30-9:25AM **Improving Question Design to Maximize Reliability and Validity**
Jon Krosnick – Stanford University

During the last 100 years, the practice of survey questionnaire design has resembled stereotypes of the Wild West: different investigators write different sorts of questionnaires with no apparent set of core principles or evidence apparently identifying best practices for maximizing reliability and validity. Yet over these years, a growing body of research, largely unrecognized by most contemporary questionnaire designers, provides guidance in making such decisions. And new such work is being published all the time. Furthermore, in many situations, researchers choose to ignore this evidence by continuing to ask old, suboptimal questions in order to permit tracking trends over time. Greater uniformity in the practice of questionnaire design and a willingness to replace suboptimal measures in new studies may be desirable.

9:25-10:20AM **Probability vs. Non-probability Methods**
Gary Langer – Langer Research Associates

Most survey researchers (including the AAPOR Task Force on Online Panels) acknowledge that probability sampling is the best way to collect the most accurate measurements describing a population of interest. However, the cost and time savings afforded by collecting data from convenience samples of volunteers has attracted many researchers to non-probability samples. Various sorts of steps have been taken to enhance the resemblance of such samples to the populations of interest, including propensity score matching, quotas, weighting, aggregation of individuals from multiple recruitment sources, and river sampling of people carrying out other activities. Understanding the effectiveness of these methods seems worthwhile for all of survey research.

10:20-10:30AM ***Coffee Break***

10:30-11:25AM **Building Household Rosters Sensibly**
Kathy Ashenfelter – U.S. Census Bureau

The vast majority of representative sample surveys begin by interviewers eliciting rosters of all household members, so a random selection can be made among them. But the rules used to determine who should be considered a household member and who should not are remarkably unstandardized across studies. For example, the U.S. Census Bureau employs different rules in different studies, and even within a single studies, different instructions to interviewers and respondents sometimes contradict one another. Thus, this is an arena in which conceptual and operational work can help survey researchers to optimize a procedure that is used across a large number of studies

11:25-12:20PM **Incentives**
Eleanor Singer – University of Michigan

This presentation summarizes what we know about the effect of incentives on various outcomes--response rates in different kinds of surveys, sample composition, response quality, and response distributions. The findings are based, whenever possible, on randomized experiments using large samples and, where possible, replicated. Since most of these experiments have been designed to improve response rates, the presentation suggests how one might think about using incentives to achieve different goals, primarily to reduce nonresponse bias. The presentation concludes with a few general "best practices" and a longer list of recommendations for needed research.

12:20-1:20PM ***Lunch Break***

1:20-2:15PM **Perception of Visual Displays and Survey Navigation**
Stephen Kosslyn – Stanford University (via video)

Surveys often make use of show cards and other visual aids for respondents yet little work has been done to bridge the understanding of visual perception developed by cognitive psychologists and best-practices in the visual design of surveys. Furthermore, despite the development of a significant empirical literature on the effects of survey design and the visual navigation of survey instruments, very little has been done in this area to bridge work from psychology on visual perception and navigation to survey methodology. The time may be ripe for interdisciplinary efforts to join survey methodologists and psychologists to conduct research aimed at developing a set of best-practices for the visual presentation of surveys and survey tools to maximize their effectiveness with respondents.

2:15-3:10PM **Data Collection Mode**
Roger Tourangeau – Westat

For several decades, beginning in the 1970s, survey data collection was dominated by three modes of data collection—face-to-face interviewing, telephone interviewing, and mail surveys. The widespread adoption of computer technology by survey researchers has multiplied the possibilities for conducting surveys. Many surveys continue to rely on face-to-face or telephone interviews but with computerized rather than paper questionnaires. However, advances in telephone technology, including the rapid proliferation of mobile telephones, has had the paradoxical effect of making telephone surveys harder to carry out than before. As a result, mail surveys sent to address-based samples may come to supplant telephone surveys for many researchers. More recently, the Internet has offered the opportunity for faster and perhaps less expensive data collection. The choice of a data collection mode has effects on all the major sources of survey error—sampling, coverage, nonresponse, and measurement error. Various features of the different mode account for their effects on survey error. These include 1) the presence of an interviewer, 2) the ability of a computerized instrument to be responsive (versus the near impossibility of responsiveness in a paper questionnaire), 3) the different demands on reading skill and computer literacy imposed by the different modes, 4) other effects on cognitive burden (such as control over the pacing of the questions in self-administered modes); and 5) the reliance on visual or aural channels to convey information. Understanding these issues can help researchers to choose among modes, to tailor the design of data collection in a particular mode, and to decide when and whether to implement multi-mode designs.

3:10-3:20PM ***Coffee Break***

3:20-4:15PM **Optimizing Response Rates**
Michael Brick – Westat

Evidence to date suggests that even low response rate surveys are remarkably robust with respect to nonresponse bias, but maximizing response rates is still a standard objective of many sample surveys. In recent years, research has generated many insights into effective practices for maximizing response rates. These goals of achieving maximal response rates have obvious benefits, but they may also have some negative consequences. This talk discusses some of the key factors that affect response rates and the potential effects of those techniques. Clearly, the field is not yet unified in approaches that can be efficient and effective.

4:15-5:10PM **Computation of Survey Weights**
Matthew DeBell – Stanford University

A standard component of survey research is the computation of weights to post-stratify survey samples and to account for intended unequal probabilities of selection due to study design. Until recently, the process of building weights was thought to be as much arts as science, adjusting a particular sample in order to maximize accuracy while minimizing the design effect. However, some large-scale infrastructure survey projects do not provide carefully constructed weights for some of their survey datasets. And although the procedures used to compute can be standardized to a large extent (e.g., by the R software package ANESRake), such standardization is not yet common across studies. For example, a public debate erupted recently when it was revealed that the Gallup Organization builds weights to match the demographic characteristics of owners of landline telephones, despite also calling cell phones. One prominent researcher went so far as to say: *"Survey weights, like sausage and legislation, are designed and best appreciated by those who are placed a respectable distance from their manufacture."*

5:10-5:15PM **Closing Remarks**
Jon Krosnick – Stanford University

The Future of Survey Research: Challenges and Opportunities

Sponsored by The National Science Foundation

November 8-9, 2012

Conference Conveners:
Jon Krosnick – *Stanford University*
Stanley Presser – *University of Maryland*
Kaye Husbands Fealing – *University of Minnesota*
Steven Ruggles – *University of Minnesota*

Conference Organizer:
David Vannette – *Stanford University*

Hosted at the Hilton Arlington
950 North Stafford Street
Arlington, Virginia 22203

STANFORD
UNIVERSITY

Conference Overview

This conference will focus on ways to augment data collection and data dissemination. The purpose here is to first describe opportunities for identifying, collecting, and linking supplemental forms of data with survey data. Presentations will describe past experiences with such linking that have been fruitful. Presentations will focus on how linkable data can be obtained, the challenges inherent in the process of linking survey data to non-survey data, and the limits of the value of data linkage.

In addition, the conference will devote time to investigating best practices in data dissemination and data collection procedure documentation by reviewing the practices of major data collection and dissemination organizations and the challenges they face.

Especially in light of recent cases of dishonesty regarding data in the social sciences, moving toward increased transparency of data collection and analysis and addressing concerns about integrity are important points to cover. It will be especially useful to outline the approaches that would be most desirable for large-scale infrastructure surveys to implement.

Conference Agenda - November 8th

7:30-8:30AM ***Continental breakfast***

8:30-8:35AM **Introduction**
Jon Krosnick – Stanford University

8:35-9:35AM **Paradata**
Frauke Kreuter – University of Maryland

Paradata are information about the process of collecting survey data in addition to the answers to the survey questions. For example, the number of seconds that it takes a respondent to answer a question is a piece of paradata that psychologists have found to be a useful analytical tool. Likewise, the time of day when an interview is conducted may be useful for understanding the responses obtained. However, the field of survey research has not endorsed or widely adopted standards for what sorts of paradata should be collected and how they should be recorded. A review of the array of potentially valuable paradata that could be collected and best practices for collecting such paradata seems worthwhile.

9:35-10:15AM **Interviewer observations**
Brady West – University of Michigan

Interviewer-administered interviews routinely end by asking interviewers to answer a series of questions about the respondent. For example, interviewers have been asked to describe how distracted the respondent was during the interview and how knowledgeable he/she seemed to be about the survey's topic. Interviewers can also make observations about the characteristics of a dwelling unit, about the respondent's neighborhood, and about other physical objects and settings. And interviewers are routinely asked to guess respondents' sexes. How accurate are these assessments? What are the best ways to design interviewer questions so as to obtain accurate information? Should the process of interviewer questioning be standardized across studies? And would interviewer training enhance the accuracy of their assessments? A small but growing body of literature has begun exploring these issues. A review of this literature and of the potential for broader use of interviewer observations seems timely.

10:15-10:25AM ***Coffee Break***

10:25-11:15AM **History and promise of blending survey data with government records on turnout**
Michael McDonald – George Mason University

Scholars have analyzed voter registration records in numerous contexts, including: to validate self-reported vote and other survey data, to measure racial voting patterns for

voting rights scholarship and litigation, to conduct voter mobilization and persuasion experiments, to analyze turnout effects of early voting laws, and to understand when individuals register to vote. A recent rise in research coincides with greater availability of data enabled by the creation of statewide voter registration databases. Yet, there has been little recent investigation into the reliability of voter registration data and how this may bias research. The records are well-known to have registrants who are no longer eligible to vote at an address, but data entry errors present potentially serious challenges, especially to research designs that rely on matching records. Better understanding the limitations of these data, and ways to overcome them, will lead to better research and have practical applications to improve election administration.

11:15-12:05PM **Challenges with validating survey data**
Matthew Berent – Stanford University

Validating survey data has provided many opportunities to evaluate the accuracy of self-report data provided by respondents. However, validation efforts are often non-trivial and present challenges at a number of different levels. While many survey projects have run validation tests when possible the amount of error still present in estimates after validation should give project managers pause before they decide to conduct a validation study. New approaches to validation may be necessary to develop if these studies are to provide their maximum benefit and the time may be ripe for exploring new approaches and techniques for validating self-report data.

12:05-1:15PM ***Lunch Break***

1:15-2:05PM **Biomarkers**
David Weir – University of Michigan

For decades, social scientists have recognized the value of collecting biological data from survey respondents to supplement self-reports. For example, in the 1970s, researchers studying cigarette smoking among children administered questionnaires and also collected saliva samples to permit analysis of thiocyanate, an indicator of tobacco use. In recent years, survey researchers have begun to collect such biological data when conducting large-scale surveys. However, this sort of data collection adds to study costs, increases required interview training, complicates interviewer work due to the necessity of preserving biological samples properly, may decrease respondent recruitment and retention, and raises ethical issues. Furthermore, some studies have shown that, to the surprise of some observers, self-reports turned out to be as accurate as or more accurate than biological measures thought initially to yield more accurate assessments of health status. Now is an excellent time to review the challenges and the potential advantages of the wide array of biological measures that might be collected to supplement conventional survey data collection.

2:05-2:55PM **Leave-behind measurement supplements**

Michael Link – The Nielsen Company

For many decades, survey researchers have supplemented data collected during face-to-face interviews with data collected via paper questionnaires that interviewers leave with respondents to be completed and mailed back to the researchers later. This sort of supplement has the advantage of expanding the poll of data collected considerably while minimizing interviewer time and allowing respondents to answer questions in small doses, taking breaks as often as they wish. This approach couples the apparent advantage of the high response rates that can be achieved by face-to-face at-home recruitment and the confidentiality afforded by self-administration. However, completion rates of self-administered leave-behind questionnaires are usually far from 100% (thereby introducing the potential for biased non-response), and some research suggests that people complete paper questionnaires quickly and superficially, rather than thinking carefully in order to answer questions accurately. The advantages and disadvantages of such measuring supplements seem worth reviewing at this time.

2:55-3:05PM ***Coffee Break***

3:05-3:55 PM **Cognitive Evaluation of Survey Instruments**
Gordon Willis – NCI/NIH

As other presentations at this Conference make clear, the world of survey research is changing rapidly, and demands fundamental changes in the way that methodologists address a range of current challenges. In addition to problems of coverage, response rates, and adjustment to technological developments, we must also pay heed to the quality of the information obtained from our survey instruments. The area of response error - and its complement, response quality -- has been a vital focus of survey methodology, especially with the advent of the Cognitive Aspects of Survey Methodology (CASM), a theory-driven interdisciplinary science which has in turn spawned applied methods for the pretesting and evaluation of survey questionnaires and other materials. In particular, Cognitive Interviewing, an intensive qualitative technique, has developed as a well-accepted approach to improving the quality of survey responses. Cognitive testing is now practiced across a range of Federal agencies, private organizations, and Universities, within an increasingly wide range of countries and contexts. At this point, however, practitioners apply a wide variety of specific approaches, lack an accepted set of best practices, conduct little objective evaluation of procedures, and have made limited efforts to critically test whether cognitive interviewing does result in a reduction in survey response error. As such, it is imperative that researchers embark on a program of systematic evaluation, in order to determine which procedural variants are efficient and effective. Given the application of relatively modest resources, it would be possible to conduct methodological research that will directly impact future practice concerning key factors such as (a) optimal sample sizes; (b) efficacy of various verbal probing techniques; and (c) appropriate data reduction and analysis techniques. Overall, research support in this area could play a large part in establishing a 'cognitive

interviewing toolbox' that appropriately matches the particular tool to the requirements of the questionnaire evaluation task to be completed.

3:55-4:45PM **Confidentiality and anonymity**
Roger Tourangeau – Westat

Surveys often collect information from respondents that are sensitive, about use of illegal drugs, cheating on income tax returns, attitudes on racial policies, sexual activities, and more. Survey researchers generally presume that assuring confidentiality and anonymity to respondents is desirable in order for respondents to be willing to participate in surveys generally and to answer sensitive questions honestly. Yet the use of Audio Computer-Assisted Self-Interviewing (whereby respondents being interviewed in their homes are given control of the interviewer's laptop computer, hear the computer read questions aloud to them privately on headphones, and answer privately by typing answers directly into the laptop) often obtain higher estimates of embarrassing attributes, suggesting that assurances of confidentiality and anonymity may not be sufficient to elicit completely honest oral responses during interviewer administered surveys. Furthermore, recent research suggests that confidentiality and anonymity may reduce the sense of accountability that respondents feel and may thereby encourage survey satisficing and compromise the accuracy of self-reports. A review of the current state of knowledge on the benefits and costs of confidentiality and anonymity seems merited.

Conference Agenda - November 9th

7:30-8:30AM ***Continental breakfast***

8:30-9:20AM **Improving information quality and availability through interactions between government, academic, and industry survey research sectors**
Robert Groves – Georgetown University

9:20-10:10AM **Linking Knowledge Networks web panel data with external data**
Josh Pasek – University of Michigan

With the move to address-based sampling, survey firms can now purchase data from commercial marketing companies to supplement information about sampled households. Because such ancillary data can provide information about non-respondents as well as respondents, it has the potential to streamline various aspects of the survey process. Among other purposes, ancillary data could be used for nonresponse adjustment, targeted sampling, and sampling frame corrections. For these potentials to be realized, however, it is important to document both the accuracy

and systematic inaccuracies of ancillary data and to assess the use of ancillary data as a corrective tool. The current study uses a unique dataset linking survey self-reports derived from an address-based sample with ancillary data collected on all sampled individuals (not just respondents). Exploring correspondence and discrepancies between the two data sources, this study identifies patterns of systematic inaccuracy and nonignorable missingness that threaten to undermine many conclusions derived with the use of consumer file data. We also use multiple imputation to produce demographic data for all individuals in the sampling frame based on the ancillary data. The results these imputations are then systematically compared with Current Population Survey estimates to assess whether matching strategies of this sort can successfully address biases due to unit non-response. We illustrate the circumstances under which the imputations produced results that were more and less representative and discuss implications for the use of such matching strategies with both probability and non-probability sampling designs. Overall the results suggest important limitations in our ability to use consumer file ancillary data for population inferences.

10:10-10:20AM ***Coffee Break***

10:20-11:10AM **Linking survey data with commercial databases**
Bob Blaemire - Catalist

What I try to do is give an overview of this business, talk about why Catalist was built and how we build the database. Then I proceed to show data that illustrates the effectiveness of voter contact using real world examples from the 2008 presidential election. I will address how academics across the country are taking advantage of the Catalist database, generally preferring actual voter turnout validation data over self-report data. Some have us take survey data and append a range of political data elements to those records, allowing them to learn lots more about the surveyed records than they otherwise would.

11:10-12:00PM **Experience sampling**
Arthur Stone – Stony Brook University

Experience sampling is a technique used mostly by psychologists, so far, to tap into people's moods, thoughts, and behaviors at randomly selected times during their days. Research participants can be signaled at random times either by a preprogrammed timing device they carry with them or by receiving a text message or a cell phone call. When prompted, participants are expected to stop their activities as quickly as possible and to answer a series of questions, which they might have on paper or could answer electronically via the Internet or a text message using a cell phone. This approach entails the advantages of obtaining reports of objective and subjective constructs in real time, without the potential inaccuracy entailed by retrospective reports, and such data can be used not only in and of themselves but also to assess the validity and reliability of self-reports. To date, researchers have yet to incorporate the collection of this sort of data into large-scale surveys. However,

participation rates may not reach 100%, and the disruption caused by regular interruptions of people's daily lives may cause experience sampling assessments to be misleading. The potential and challenges of blending experience sampling with survey data collection seem worth reviewing.

12:00-1:00PM ***Lunch Break***

1:00-1:50PM **Linking survey data to official government records**
Joseph Sakshaug – Ludwig Maximilians University of Munich

Linking respondents' answers with other data available on those same people from other data sources may enhance the analytic potential of survey data. For example, some state government agencies routinely release data files that list all residents who were registered to vote in an election and indicate whether they did vote. Other government agencies also maintain elaborate official records on individuals, including records on income taxes (maintained by the Internal Revenue Service), on military activities, on receipt of Medicare, and much more. Census Research Data Centers have been created across the U.S. to permit such data linking, and commercial companies such as Catalyst offer linkage services. The range of analyses possible for a researcher would widen vastly with such additional records. However, accurately linking survey responses to official records requires the collection of detailed and potentially sensitive information from respondents (e.g., social security numbers in addition to full names and addresses), which may lead to increased non-response. The process of record linkage can be also be inaccurate and therefore misleading. Some commercial companies insist on keeping their matching procedures confidential thereby preventing assessment of accuracy. The advantages, challenges, and dangers of record linkage seem worth exploring.

1:50-2:40PM **Research Transparency and the Credibility of Survey-Based Social Science**
Arthur Lupia – University of Michigan

Across a number of disciplines, scholars improperly interpret or are unable to replicate highly visible survey-based claims. Collectively, these failures undermine the aggregate credibility of survey research. I will describe constructive steps that researchers and organizations can take to increase the credibility of many survey-based endeavors. A core focus of my presentation is on how to improve research transparency. By research transparency, I mean documenting processes and decisions, and actions that convert labor and capital first into survey data and then into survey-based empirical claims. Research transparency has two components: production transparency and analytic transparency. Production transparency implies providing information about the procedures that converts labor and capital into data points. Today, many surveys provide little to no information about critical procedures. When users cannot access this information, their ability to correctly interpret survey data

can be severely inhibited. We have found multiple instances of scholars misinterpreting ANES data for this reason. Analytic Transparency implies a full account of how a researcher drew conclusions from a given survey, clearly mapping the path from the data to empirical claims. We have found many instances where such paths are neither public nor reproducible. I conclude by showing how meeting a small number of transparency requirements at each stage of a research process can help scholars can protect their own credibility and contribute to the credibility and perceived legitimacy of survey research generally.

2:40-2:50PM ***Coffee Break***

2:50-3:40PM **Meta data and preservation**
Steve Ruggles – University of Minnesota (via video)

Many millions of dollars have been invested by funding agencies to support the collection of survey data by many organizations for many decades. Such data are made available to potential analysts in two primary ways: (1) through archives such as the *Inter-University Consortium for Political and Social Research* and the Roper Center for Public Opinion Research, and (2) via websites maintained by the central offices of the long-term projects that collected the data, such as the National Longitudinal Surveys of Youth, The Survey of Consumers, and the National Longitudinal Study of Adolescent Health. There is remarkably little standardization of the nature of information made available to users and the formats in which this information is made available. For example, questionnaires are available for some surveys and not others. Likewise, show cards are available for some surveys and not others. Project directors often make improvisational decisions about what information to retain and disseminate rather than relying on a set of conventions or best practices. Investing the necessary time and effort into developing a set of best practices may be useful.

3:40-4:30PM **Specialized tools for measuring past events**
Robert Belli – University of Nebraska-Lincoln

Survey researchers often seek to document respondents' past experiences, such as the dates a person worked for pay, or the extent of television viewing. In recent decades, specialized tools have been developed in order to assist respondents in recalling their pasts and to facilitate accurate measurement. For example, calendar interviewing has been used to help respondents reconstruct decades of experiences, and time use surveys have made use of diaries for tracking respondent behaviors over a single day. These and other data collection tools offer opportunities to collect better quality data from respondents in comparison to estimates derived from conventional standardized interviews, as they more readily tap into cues existing in the structure of autobiographical memory. Additional methodological work with recorded interviews

and paradata derived from calendars and time diaries is needed to understand the interactional processes among interviewers, respondents, and instruments that enhance data quality. This work will enable the development of interviewer- and self-administered computerized smart instruments

4:30-5:20PM **Usability of survey project websites**
David L. Vannette – Stanford University

As users of survey data have increasingly come to depend on the Internet to access these data, wide variation has developed between the interfaces employed to make such data available. Furthermore, dataset formats and website designs are often strikingly different across projects and archives, so a user must navigate the idiosyncrasies of each dissemination platform in order to obtained desired information. We conducted a number of usability tests on both project websites and archive websites to identify and evaluate the most effective approaches to organizing and disseminating data on the web. Design and structural principles discussed here may be useful for project directors and archive managers that want to make their web platforms increasingly user-friendly.

5:20-5:30PM **Closing remarks**
Jon Krosnick – Stanford University